A&A 649, A6 (2021)
https://doi.org/10.1051/0004-6361/202039498

Astronomy & Astrophysics

Special issue

Gaia Early Data Release 3

Gaia Early Data Release 3

The *Gaia* Catalogue of Nearby Stars★

Gaia Collaboration: R. L. Smart[1,★★], L. M. Sarro[2], J. Rybizki[3], C. Reylé[4], A. C. Robin[4], N. C. Hambly[5], U. Abbas[1], M. A. Barstow[6], J. H. J. de Bruijne[7], B. Bucciarelli[1], J. M. Carrasco[8], W. J. Cooper[9,1], S. T. Hodgkin[10], E. Masana[8], D. Michalik[7], J. Sahlmann[11], A. Sozzetti[1], A. G. A. Brown[12], A. Vallenari[13], T. Prusti[7], C. Babusiaux[14,15], M. Biermann[16], O. L. Creevey[17], D. W. Evans[10], L. Eyer[18], A. Hutton[19], F. Jansen[7], C. Jordi[8], S. A. Klioner[20], U. Lammers[21], L. Lindegren[22], X. Luri[8], F. Mignard[17], C. Panem[23], D. Pourbaix[24,25], S. Randich[26], P. Sartoretti[15], C. Soubiran[27], N. A. Walton[10], F. Arenou[15], C. A. L. Bailer-Jones[3], U. Bastian[16], M. Cropper[28], R. Drimmel[1], D. Katz[15], M. G. Lattanzi[1,29], F. van Leeuwen[10], J. Bakker[21], J. Castañeda[30], F. De Angeli[10], C. Ducourant[27], C. Fabricius[8], M. Fouesneau[3], Y. Frémat[31], R. Guerra[21], A. Guerrier[23], J. Guiraud[23], A. Jean-Antoine Piccolo[23], R. Messineo[32], N. Mowlavi[18], C. Nicolas[23], K. Nienartowicz[33,34], F. Pailler[23], P. Panuzzo[15], F. Riclet[23], W. Roux[23], G. M. Seabroke[28], R. Sordo[13], P. Tanga[17], F. Thévenin[17], G. Gracia-Abril[35,16], J. Portell[8], D. Teyssier[36], M. Altmann[16,37], R. Andrae[3], I. Bellas-Velidis[38], K. Benson[28], J. Berthier[39], R. Blomme[31], E. Brugaletta[40], P. W. Burgess[10], G. Busso[10], B. Carry[17], A. Cellino[1], N. Cheek[41], G. Clementini[42], Y. Damerdji[43,44], M. Davidson[5], L. Delchambre[43], A. Dell'Oro[26], J. Fernández-Hernández[45], L. Galluccio[17], P. García-Lario[21], M. Garcia-Reinaldos[21], J. González-Núñez[41,46], E. Gosset[43,25], R. Haigron[15], J.-L. Halbwachs[47], D. L. Harrison[10,48], D. Hatzidimitriou[49], U. Heiter[50], J. Hernández[21], D. Hestroffer[39], B. Holl[18,33], K. Janßen[51], G. Jevardat de Fombelle[18], S. Jordan[16], A. Krone-Martins[52,53], A. C. Lanzafame[40,54], W. Löffler[16], A. Lorca[19], M. Manteiga[55], O. Marchal[47], P. M. Marrese[56,57], A. Moitinho[52], A. Mora[19], K. Muinonen[58,59], P. Osborne[10], E. Pancino[26,57], T. Pauwels[31], A. Recio-Blanco[17], P. J. Richards[60], M. Riello[10], L. Rimoldini[33], T. Roegiers[61], C. Siopis[24], M. Smith[28], A. Ulla[62], E. Utrilla[19], M. van Leeuwen[10], W. van Reeven[19], A. Abreu Aramburu[45], S. Accart[63], C. Aerts[64,65,3], J. J. Aguado[2], M. Ajaj[15], G. Altavilla[56,57], M. A. Álvarez[66], J. Álvarez Cid-Fuentes[67], J. Alves[68], R. I. Anderson[69], E. Anglada Varela[45], T. Antoja[8], M. Audard[33], D. Baines[36], S. G. Baker[28], L. Balaguer-Núñez[8], E. Balbinot[70], Z. Balog[16,3], C. Barache[37], D. Barbato[18,1], M. Barros[52], S. Bartolomé[8], J.-L. Bassilana[63], N. Bauchet[39], A. Baudesson-Stella[63], U. Becciani[40], M. Bellazzini[42], M. Bernet[8], S. Bertone[71,72,1], L. Bianchi[73], S. Blanco-Cuaresma[74], T. Boch[47], A. Bombrun[75], D. Bossini[76], S. Bouquillon[37], A. Bragaglia[42], L. Bramante[32], E. Breedt[10], A. Bressan[77], N. Brouillet[27], A. Burlacu[78], D. Busonero[1], A. G. Butkevich[1], R. Buzzi[1], E. Caffau[15], R. Cancelliere[79], H. Cánovas[19], T. Cantat-Gaudin[8], R. Carballo[80], T. Carlucci[37], M. I Carnerero[1], L. Casamiquela[27], M. Castellani[56], A. Castro-Ginard[8], P. Castro Sampol[8], L. Chaoul[23], P. Charlot[27], L. Chemin[81], A. Chiavassa[17], M.-R. L. Cioni[51], G. Comoretto[82], T. Cornez[63], S. Cowell[10], F. Crifo[15], M. Crosta[1], C. Crowley[75], C. Dafonte[66], A. Dapergolas[38], M. David[83], P. David[39], P. de Laverny[17], F. De Luise[84], R. De March[32], J. De Ridder[64], R. de Souza[85], P. de Teodoro[21], A. de Torres[75], E. F. del Peloso[16], E. del Pozo[19], A. Delgado[10], H. E. Delgado[2], J.-B. Delisle[18], P. Di Matteo[15], S. Diakite[86], C. Diener[10], E. Distefano[40], C. Dolding[28], D. Eappachen[87,65], B. Edvardsson[88], H. Enke[51], P. Esquej[11], C. Fabre[89], M. Fabrizio[56,57], S. Faigler[90], G. Fedorets[58,91], P. Fernique[47,92], A. Fienga[93,39], F. Figueras[8], C. Fouron[78], F. Fragkoudi[94], E. Fraile[11], F. Franke[95], M. Gai[1], D. Garabato[66], A. Garcia-Gutierrez[8], M. García-Torres[96], A. Garofalo[42], P. Gavras[11], E. Gerlach[20], R. Geyer[20], P. Giacobbe[1], G. Gilmore[10], S. Girona[67], G. Giuffrida[56], R. Gomel[90], A. Gomez[66], I. Gonzalez-Santamaria[66], J. J. González-Vidal[8], M. Granvik[58,97], R. Gutiérrez-Sánchez[36], L. P. Guy[33,82], M. Hauser[3,98], M. Haywood[15], A. Helmi[70], S. L. Hidalgo[99,100], T. Hilger[20], N. Hładczuk[21], D. Hobbs[22], G. Holland[10], H. E. Huckle[28], G. Jasniewicz[101], P. G. Jonker[65,87], J. Juaristi Campillo[16], F. Julbe[8], L. Karbevska[18], P. Kervella[102], S. Khanna[70], A. Kochoska[103], M. Kontizas[49], G. Kordopatis[17], A. J. Korn[50], Z. Kostrzewa-Rutkowska[12,87], K. Kruszyńska[104], S. Lambert[37], A. F. Lanza[40], Y. Lasne[63], J.-F. Le Campion[105], Y. Le Fustec[78], Y. Lebreton[102,106], T. Lebzelter[68], S. Leccia[107], N. Leclerc[15], I. Lecoeur-Taibi[33], S. Liao[1], E. Licata[1], H. E. P. Lindstrøm[1,108], T. A. Lister[109], E. Livanou[49], A. Lobel[31], P. Madrero Pardo[8], S. Managau[63], R. G. Mann[5], J. M. Marchant[110], M. Marconi[107], M. M. S. Marcos Santos[41], S. Marinoni[56,57], F. Marocco[111,112], D. J. Marshall[113], L. Martin Polo[41], J. M. Martín-Fleitas[19], A. Masip[8], D. Massari[42], A. Mastrobuono-Battisti[22], T. Mazeh[90],

★ Tables are only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsarc.u-strasbg.fr/viz-bin/cat/J/A+A/649/A6

★★ Corresponding author; e-mail: richard.smart@inaf.it

P. J. McMillan[22], S. Messina[40], N. R. Millar[10], A. Mints[51], D. Molina[8], R. Molinaro[107], L. Molnár[114,115,116], P. Montegriffo[42], R. Mor[8], R. Morbidelli[1], T. Morel[43], D. Morris[5], A. F. Mulone[32], D. Munoz[63], T. Muraveva[42], C. P. Murphy[21], I. Musella[107], L. Noval[63], C. Ordénovic[17], G. Orrù[32], J. Osinde[11], C. Pagani[6], I. Pagano[40], L. Palaversa[117,10], P. A. Palicio[17], A. Panahi[90], M. Pawlak[118,104], X. Peñalosa Esteller[8], A. Penttilä[58], A. M. Piersimoni[84], F.-X. Pineau[47], E. Plachy[114,115,116], G. Plum[15], E. Poggio[1], E. Poretti[119], E. Poujoulet[120], A. Prša[103], L. Pulone[56], E. Racero[41,121], S. Ragaini[42], M. Rainer[26], C. M. Raiteri[1], N. Rambaux[39], P. Ramos[8], M. Ramos-Lerate[122], P. Re Fiorentin[1], S. Regibo[64], V. Ripepi[107], A. Riva[1], G. Rixon[10], N. Robichon[15], C. Robin[63], M. Roelens[18], L. Rohrbasser[33], M. Romero-Gómez[8], N. Rowell[5], F. Royer[15], K. A. Rybicki[104], G. Sadowski[24], A. Sagristà Sellés[16], J. Salgado[36], E. Salguero[45], N. Samaras[31], V. Sanchez Gimenez[8], N. Sanna[26], R. Santoveña[66], M. Sarasso[1], M. Schultheis[17], E. Sciacca[40], M. Segol[95], J. C. Segovia[41], D. Ségransan[18], D. Semeux[89], S. Shahaf[90], H. I. Siddiqui[123], A. Siebert[47,92], L. Siltala[58], E. Slezak[17], E. Solano[124], F. Solitro[32], D. Souami[102,125], J. Souchay[37], A. Spagna[1], F. Spoto[74], I. A. Steele[110], H. Steidelmüller[20], C. A. Stephenson[36], M. Süveges[33,126,3], L. Szabados[114], E. Szegedi-Elek[114], F. Taris[37], G. Tauran[63], M. B. Taylor[127], R. Teixeira[85], W. Thuillot[39], N. Tonello[67], F. Torra[30], J. Torra[†,8], C. Turon[15], N. Unger[18], M. Vaillant[63], E. van Dillen[95], O. Vanel[15], A. Vecchiato[1], Y. Viala[15], D. Vicente[67], S. Voutsinas[5], M. Weiler[8], T. Wevers[10], Ł. Wyrzykowski[104], A. Yoldas[10], P. Yvard[95], H. Zhao[17], J. Zorec[128], S. Zucker[129], C. Zurbach[130], and T. Zwitter[131]

(Affiliations can be found after the references)

Received 22 September 2020 / Accepted 30 October 2020

ABSTRACT

Aims. We produce a clean and well-characterised catalogue of objects within 100 pc of the Sun from the *Gaia* Early Data Release 3. We characterise the catalogue through comparisons to the full data release, external catalogues, and simulations. We carry out a first analysis of the science that is possible with this sample to demonstrate its potential and best practices for its use.

Methods. The selection of objects within 100 pc from the full catalogue used selected training sets, machine-learning procedures, astrometric quantities, and solution quality indicators to determine a probability that the astrometric solution is reliable. The training set construction exploited the astrometric data, quality flags, and external photometry. For all candidates we calculated distance posterior probability densities using Bayesian procedures and mock catalogues to define priors. Any object with reliable astrometry and a non-zero probability of being within 100 pc is included in the catalogue.

Results. We have produced a catalogue of 331 312 objects that we estimate contains at least 92% of stars of stellar type M9 within 100 pc of the Sun. We estimate that 9% of the stars in this catalogue probably lie outside 100 pc, but when the distance probability function is used, a correct treatment of this contamination is possible. We produced luminosity functions with a high signal-to-noise ratio for the main-sequence stars, giants, and white dwarfs. We examined in detail the Hyades cluster, the white dwarf population, and wide-binary systems and produced candidate lists for all three samples. We detected local manifestations of several streams, superclusters, and halo objects, in which we identified 12 members of *Gaia* Enceladus. We present the first direct parallaxes of five objects in multiple systems within 10 pc of the Sun.

Conclusions. We provide the community with a large, well-characterised catalogue of objects in the solar neighbourhood. This is a primary benchmark for measuring and understanding fundamental parameters and descriptive functions in astronomy.

Key words. catalogs – astrometry – stars: luminosity function, mass function – Hertzsprung-Russell and C-M diagrams – stars: low-mass – solar neighborhood

1. Introduction

The history of astronomical research is rich with instances in which improvements in our observational knowledge have led to breakthroughs in our theoretical understanding. The protracted astronomical timescales have required astronomers to employ significant ingenuity to extrapolate today's snapshot in time to understanding the history and evolution of even the local part of our Galaxy. This is hampered by the fact that our knowledge and census of the Galaxy, including the local region, is incomplete. The difficulty has primarily been in the resources required to determine distances and the lack of a sufficiently deep and complete census of nearby objects, both of which will be resolved by the ESA *Gaia* mission. *Gaia* will determine distances, motions, and colours of all the stars, except for the very brightest, in the solar neighbourhood.

The solar neighbourhood has been considerably studied since the beginning of the past century when astronomers began to routinely measure stellar parallaxes. In 1957 this effort was formalised with the publication of 915 known stars within 20 pc (Gliese 1957). Various updates and extensions to larger distances produced what became the Catalogue of Nearby Stars, including all known stars, 3803, within 25 pc released in 1991 (CNS, Gliese & Jahreiß 1991). The HIPPARCOS mission increased the quantity and quality of the CNS content; however, the magnitude limit of HIPPARCOS resulted in an incompleteness for faint objects. In 1998 the CNS dataset was moved online[1] and currently has 5835 entries, but it is no longer updated. The most recent update of the CNS by Stauffer et al. (2010) was to provide accurate coordinates and near-infrared magnitudes taken from the Two Micron Sky Survey (2MASS, Skrutskie et al. 2006).

The CNS has been used in various investigations, gathering over 300 citations from the studies of wide-binary systems (Caballero 2010; Lowrance et al. 2002; Poveda et al. 1994;

[†] Deceased.

[1] https://wwwadd.zah.uni-heidelberg.de/datenbanken/aricns/

Latham et al. 1991), searches for solar twins (Friel et al. 1993), statistics for extra-solar planet hosts (Biller et al. 2007; Johnson et al. 2007; Pravdo et al. 2006), the local luminosity function (Reid et al. 2002; Gizis & Reid 1999; Martini & Osmer 1998; Wielen et al. 1983; Reid & Gizis 1997), the mass-luminosity relation (Henry et al. 1999), to galactic and local kinematics (Bienayme & Sechaud 1997; Wielen 1974). The utility of the CNS has been limited by its incompleteness and the lack of high-precision parallaxes. Other compilations of nearby objects have either limited the type of objects to, for example, ultra-cool dwarfs and 25 pc (Bardalez Gagliuffi et al. 2019), cooler T/Y dwarfs and 20 pc (Kirkpatrick et al. 2019), complete spectral coverage but limited volume, such as the REsearch Consortium On Nearby Stars 10 pc sample (Henry et al. 2018), or, with the inclusion of substellar objects and an 8 pc volume (Kirkpatrick et al. 2012). However, these catalogues have by necessity all been based on multiple observational sources and astrometry of limited precision. The high astrometric precision and faint magnitude survey mode of *Gaia* will provide a census that will be more complete, in a larger volume, and homogeneous. It is therefore easier to characterise.

In this contribution we present the *Gaia* Catalogue of Nearby Stars (GCNS), a first attempt to make a census of all stars in the solar neighbourhood using the *Gaia* results. In the GCNS we define the solar neighbourhood to be a sphere of radius 100 pc centred on the Sun. This will be volume-complete for all objects earlier than M8 at the nominal $G = 20.7$ magnitude limit of *Gaia*. Later type objects will be too faint for *Gaia* at 100 pc, resulting in progressively smaller complete volumes with increasing spectral type. In Sect. 2 we discuss the generation of the GCNS, in Sect. 3 we present an overview of the catalogue contents and availability, in Sect. 4 we carry out some quality assurance tests, and in Sect. 5 we report an example for a scientific exploitation of the GCNS.

2. GCNS generation

In this section we describe the process by which we have generated the GCNS starting from a selection of all sources in the *Gaia* EDR3 archive with measured parallaxes $\hat{\varpi} > 8$ mas (we use ϖ for true parallaxes and $\hat{\varpi}$ for measured parallaxes). The process is composed of two phases: in the first phase (Sect. 2.1), we attempt to remove sources with spurious astrometric solutions using a random forest classifier (Breiman 2001); and in the second phase (Sect. 2.2), we infer posterior probability densities for the true distance of each source. The GCNS is then defined based on the classifier probabilities and the properties of the distance posterior distribution according to criteria specified below. These procedures are critical for the catalogue generation, and the details pertain to the area of machine-learning.

2.1. Removal of spurious sources

In order to generate the first selection of sources inside 100 pc, we constructed a classifier to identify poor astrometric solutions that result in observed parallaxes greater than 10 mas from true sources within the 100 pc radius. For objects with *Gaia*, $G = 20$, the median uncertainty of *Gaia* EDR3 parallaxes is 0.5 mas (Seabroke et al. 2021) and the global zero-point is between -20 and $-40\,\mu$as (Lindegren et al. 2021a), therefore the 10 mas boundary is extremely well defined. We started by selecting a sample with $\hat{\varpi} \geq 8$ mas to minimise the sample size and avoid introducing a large loss of sources due to the parallax measurement uncertainty. Using the GeDR3mock catalogue (Rybizki et al. 2020, cf. Sect. 2.2), we estimate that about 55 sources lie truly within 100 pc but are lost in the primary selection at 8 mas. We find a total of 121 1740 sources with measured parallaxes $\hat{\varpi} \geq 8$ mas.

Spurious astrometric solutions can be due to a number of reasons, but the causes that produce such large parallaxes are mostly related to the inclusion of outliers in the measured positions because close pairs are only resolved for certain transits and scan directions (see Sect. 7.9 of Gaia Collaboration 2018b). This is more likely to occur in regions of high surface density of sources or for close binary systems (either real or due to perspective effects). Parallax errors of smaller magnitude are more likely due to the presence of more than one object in the astrometric window or to binary orbital motion that is not accounted for.

We aim at classifying sources into two categories based solely on astrometric quantity and quality indicators. We explicitly leave photometric measurements out of the selection in order to avoid biases from preconceptions relative to the loci in the colour-absolute magnitude diagram (CAMD) where sources are expected. A classifier that uses the position of sources in the CAMD, and is therefore trained with examples from certain regions in this diagram, such as the main sequence, red clump, or white dwarf (hereafter WD) sequences, might yield an incomplete biased catalogue in the sense that sources out of these classical loci would be taken for poor astrometric solutions. In contrast, we aim at separating the two categories (loosely speaking, good and poor astrometric solutions) based on predictive variables other than those arising from the photometric measurements, and use the resulting CAMD as external checks of the selection procedure. This will allow us to identify true nearby objects with problematic photometry, as we show in subsequent sections.

In order to construct the classification model, we created a training set with examples in both categories as follows. For the set of poor astrometric solutions, we queried the *Gaia* EDR3 archive for sources with parallaxes $\hat{\varpi} < -8$ mas. The query returned 512 288 sources. We assumed that the mechanism by which large (in absolute value) spurious parallaxes are produced is the same regardless of the sign and that the distribution of astrometric quantities that the model infers from this set of large negative parallaxes is therefore equivalent (i.e. unbiased with respect) to that of the set of large spurious parallaxes. We include in Appendix A.2 a series of histograms with the distributions of the predictive variables in both the training set and the resulting classification. The latter is inevitably a consequence of the former (the training set), but the good match of the distributions for the $\hat{\varpi} < -8$ mas (training set) and $\hat{\varpi} > 8$ mas (sources classified as poor astrometric solutions) is reassuring.

Sources with poor astrometric solutions are expected to have small true parallaxes (we estimate their mean true parallax to be 0.25 mas, as justified below) and are scattered towards high absolute values due to data reduction problems, as those described above. By using the large negative parallax sample as training set for the class of poor astrometric solutions, we avoided potential contamination by sources that lie truly within the 125 pc radius or the incompleteness (and therefore bias) associated with the selection of only very clear cases of poor astrometry.

The set of examples of good astrometric solutions within the 8 mas limit was constructed as follows. We first selected sources in low-density regions of the sky (those with absolute values of the Galactic latitudes greater than 25° and at angular distances from the centres of the Large and Small Magellanic Clouds greater than 12 and 9 degrees, respectively) and kept only

sources with a positive cross-match in the 2MASS catalogue. As a result, we assembled a set of 291 030 sources with photometry in five bands: G, $G_{\rm RP}$, J, H, and K. We avoided the use of $G_{\rm BP}$ magnitudes because they have known limits for faint red objects (see Sect. 8 of Riello et al. 2021).

From these we constructed a representation space with one colour index ($G - J$) and four absolute magnitudes ($M_G, M_{\rm RP}, M_H$, and M_K). We fit models of the source distribution in the loci of WDs, the red clump and giant branch, and the main sequence. The models for the WDs, giant branch, and red clump stars are Gaussian mixture models, while the main-sequence model is based on the 5D principal curve (Hastie & Stuetzle 1989). We used these models to reject sources with positions in representation space far from these high-density loci (presumably due to incorrect cross-matches or poor astrometry). As a result, we obtained a set of 274 108 sources with consistent photometry in the *Gaia* and 2MASS bands. This is less than half the number of sources with parallaxes more negative than -8 mas. We recall that the selection of this set of examples of good astrometric solutions is based on photometric measurements and parallaxes, but we only required that the photometry in the five bands is consistent. The photometric information is not used later on, and the subsequent classification of all sources into the two categories of good and spurious astrometric measurements is based only on the astrometric quantities described below. This selection would therefore only bias the resulting catalogue if it excluded sources with good astrometric solutions whose astrometric properties were significantly different from those of the training examples.

The classification model consists of a random forest (Breiman 2001) trained on predictor variables selected from a set of 41 astrometric features listed in Table A.1. Table A.1 includes the feature names as found in the *Gaia* archive and its importance measured with the mean decrease in accuracy (Breiman 2002, two leftmost columns) or Gini index (Gini 1912, two rightmost columns). We selected features (based on the Gini index) even though random forests inherently down-weight the effect of unimportant features. We did this for the sake of efficiency. The selected features are shaded in grey in Table A.1, and we shade in red one particular variable (`astrometric_params_solved`) that can only take two values and was not selected despite the nominal relevance. The set of $2 \times 274\,108$ examples (we selected exactly the same number of examples in the two categories and verify the validity of this balanced training set choice below) was divided into a training set (67%) and a test set (33%) in order to assess the accuracy of the classifier and determine the probability threshold that optimises completeness and contamination. We find the optimum probability in the corresponding receiver operating curve (ROC), which is $p = 0.38$, yielding a sensitivity of 0.9986 (the fraction of correctly classified good examples in the test set) and a specificity of 0.9991 (the same fraction, but for the poor category). The random forest consists of 5000 decision trees built by selecting amongst three randomly selected predictors at each split. Variations in the number of trees or candidate predictors did not produce better results, as evaluated on the test set. These can be summarised by the confusion matrix shown in Table 1.

Figure 1 shows the distribution in the sky of selected (top) and rejected (bottom) sources. The distribution of selected sources looks uniform, as expected, with the exception of the slight over-density at $l, b \approx (300, 10)$ that is probably part of the Lower Centaurus Crux subgroup of the Sco OB2 association at 115 pc (Zari et al. 2018). The bottom panel highlights problematic sky areas related to high surface density regions

Table 1. Confusion matrix of the classifier evaluated in the test set.

	1	2
1	89706	128
2	83	90119

Notes. Class 1 represents good astrometric solutions (positives), and class 2 represents poor solutions (negatives). The first row shows the number of class 1 examples classified as good astrometric solutions (true positives, first column) and as poor solutions (false negatives, second column). The second row shows the number of class 2 examples classified as class 1 (false positives, first column) and class 2 (true negatives; second columns). The total number of misclassifications for the set of test examples is 0.1%.

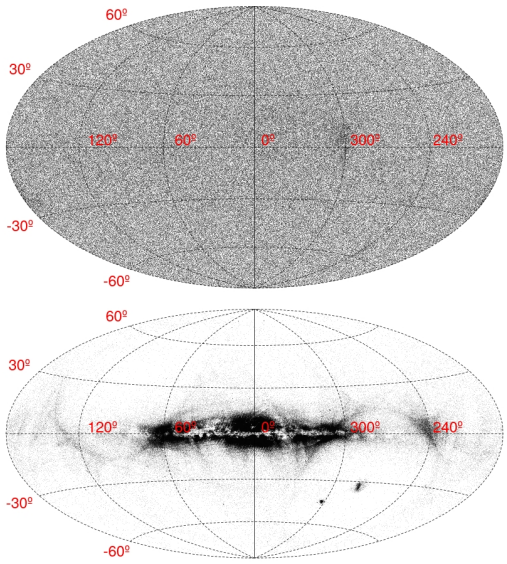

Fig. 1. Distribution of selected (*top panel*) and rejected (*bottom panel*) sources according to the random forest classifier in Galactic coordinates in an Aitoff projection.

and/or specificities of the scanning law. In order to detect signs of incompleteness and/or contamination, we inspected the distribution of absolute G magnitudes for both sets of sources (Fig. 2 left panel). The distribution of spurious sources shows a main component centred at $M_G \sim 15$; this coincides with a local bump in the distribution function of selected sources, which may be indicative of contamination.

Figure 3 shows a (logarithmic) histogram with the derived membership probabilities. Neither the number of sources with $\hat{\varpi} < -8$ mas or the comparison of the numbers of sources classified as good and poor provide evidence for a significant imbalance in the true proportions of the classes. We therefore discarded the revision of the training set proportions or the inclusion of additional actions to recalibrate the classification probabilities due to a class imbalance. Finally, the right panel of Fig. 2 shows a colour-absolute magnitude diagram (CAMD) for the full sample colour-coded by probability p. The rejected sources are predominantly in areas of the CAMD that are usually empty, consistent with our hypothesis that the parallaxes are unreliable.

As final confirmation for the assumptions underlying the training set definition we attempted to estimate the mean true parallax of the poor astrometric solution by determining the negative value of the observed parallax that results in approximately the same number of sources as those classified as poor astrometric solutions by our random forest classifier. We find 638 796 sources classified as poor astrometric solutions, which is similar

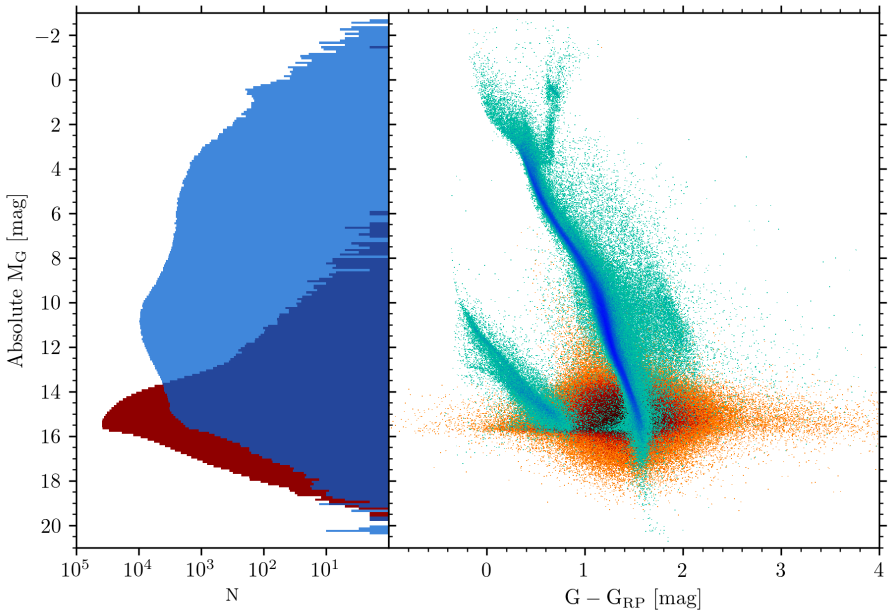

Fig. 2. *Left panel*: distribution of absolute G magnitudes for the full *Gaia* EDR3 $\hat{\varpi} \geq 8$ mas sample. The blue distribution is for selected sources and the red one for rejected sources using a bin size of $\sigma_{M_G} = 0.1$ mag. The slight bump in the distribution of selected sources at $M_G = 15$ mag that coincides with the maximum of the rejected sources is probably indicative of contamination. *Right panel*: CAMD diagram for the full sample. The blue points are good solutions and the red poor ones. The strip of source with nominally good solutions connecting the main and white dwarf sequence at $M_G \sim 15$ is unexpected and due to contamination of the GCNS by faint objects at distances of 80–120 pc, as discussed in Sect. 4.5.

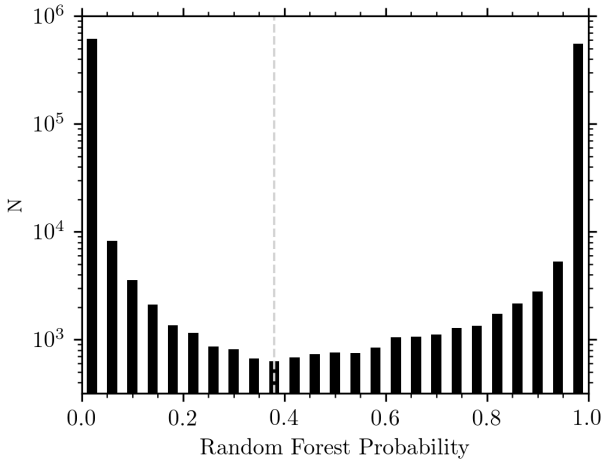

Fig. 3. Histogram of the classification probabilities (in the category of good astrometric solutions) produced by the random forest. The vertical axis is in logarithmic scale.

to the number of sources with $\hat{\varpi} \leq -7.5$ mas (639 058). If the distribution of true parallaxes of sources with poor astrometric solutions were symmetric (which is not necessarily true), then its mode could be estimated as $(8 - 7.5)/2 = 0.25$ mas or 4 kpc.

The random forest classifier described above is a solution for the particular problem of separating good and poor astrometric solutions in the solar neighbourhood, but it is not applicable at larger distances. Good and poor astrometric solutions are well separated in the space of input variables because the former are of exquisite quality. As the measured parallax decreases, the proportions of both classes change in the input parameter space and the degree of overlap between the two increases. We therefore expect misclassifications to increase for smaller observed parallaxes, also because the fraction of sources in each category varies and increases more steeply for the poor astrometric solutions. Finally, we would like to emphasise that a probability below the selection threshold does not necessarily mean that the source does not lie within 100 pc. The astrometric solution of a source can be problematic (and the source therefore rejected by the random forest) even if it is located within 100 pc.

2.2. Simple bayesian distance estimation

In order to infer distances from the observed parallaxes, we need an expected distance distribution (prior) for the sources in our sample selection ($\hat{\varpi} \geq 8$ mas). We assumed that we have removed all poor solutions. The simplest prior is a single distribution that does not depend on sky position or type of star (e.g. colour). We defined an empirical prior based on synthetic samples using the GeDR3mock, which includes all the stars down to $G = 20.7$ mag. The parallax uncertainty for GeDR3mock was empirically trained on *Gaia* DR2 data and was lowered according to the longer time baseline of *Gaia* EDR3. The mock `parallax_error` distribution is narrower than that of the empirical *Gaia* EDR3, therefore we artificially increased the spread in log(`parallax_error`), see the query below. Because the catalogue only contains the true parallaxes, we selected observed parallaxes through the following query, which can be performed on the GAVO TAP service[2]:

```
SELECT * FROM(
SELECT parallax, GAVO_RANDOM_NORMAL(parallax,
    POWER(10, ((LOG10(parallax_error)+1)*1.3)-1)) AS
    parallax_obs
-- This adds observational noise to the true
    parallaxes
FROM gedr3mock.main) AS sample
WHERE parallax_obs > 8
```

This retrieves a catalogue with 762 230 stars[3]. Their underlying true distance distribution is shown in Fig. 4. The distribution of mock stars was inspected by comparing an in-plane, $|b| < 5°$, and an out-of-plane, $|b| > 65°$. We found a 15% deficiency of stars at 100 pc distance for the out-of-plane sample, as expected due to the stratification in the z direction. When selecting for specific stellar types, the directional dependence can increase further, for instance for dynamically cold stellar populations. Here we ignored these possibilities and used a distance prior

[2] http://dc.g-vo.org/tap

[3] The number of stars retrieved will slightly change each time the query is run because the random number generator does not accept seeds.

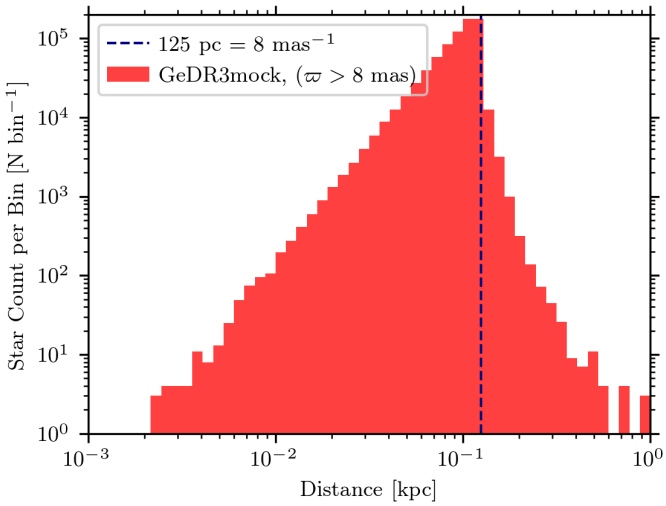

Fig. 4. Distance distribution of stars in GeDR3mock selected on observed parallax >8 mas. We use this distribution as a prior for our simple Bayesian distance estimation.

independent of colour or direction in the sky to let the exquisite data speak for themselves.

We sampled the posterior probability density function (PDF) using Markov chain Monte Carlo methods (Foreman-Mackey et al. 2013). The reported values, included in a table available at the CDS, are the percentiles (from 1 to 99) of the stabilised chain, that is, `dist_50` represents the median of the posterior distance estimation and `dist_16`, `dist_84` the lower and upper 1σ uncertainties. We also report `mean_acceptance_fractions` and `mean_autocorrelation_time` as quality indicators (but not all sources have the latter).

3. The *Gaia* catalogue of nearby stars

We now discuss the selection from the 1 211 740 objects with $\hat{\varpi} > 8$ mas for inclusion in the GCNS. As indicated in Sect. 2.1, the optimal probability threshold indicated by the ROC is $p = 0.38$. To enable a correct use of the distance PDF produced in Sect. 2.2, we retain all entries with a non-zero probability of being inside 100 pc, for which we used the distance with 1% probability, `dist_1`.

Therefore the selection for inclusion in the GCNS is:

$$p > = 0.38 \;\&\&\; \texttt{dist\_1} < = 0.1 \text{ Kpc}. \quad (1)$$

This selection resulted in 331 312 objects that are listed in the table available at the CDS, an example of which is reported in Table 2. The 880 428 objects from the full $\hat{\varpi} > 8$ mas that did not meet these criteria are provided in an identical table should they be needed for characterisation[4].

Our goal is to provide a stand-alone catalogue that will be useful when observing or for simple exploratory studies. Following this goal, we have retained minimal *Gaia* EDR3 information, source ID, basic astrometry, photometry, and a few of the quality flags used in this paper. In keeping with the *Gaia* data release policy, we do not provide uncertainties on the magnitudes but the `mean_flux_over_error` for each passband. There are 3016 objects in the full $\hat{\varpi} > 8$ mas sample that do not have *Gaia* G magnitudes, 431 of which meet our selection criteria. These 431 objects have on-board estimates of the G magnitude in the 11–13 range, and we refer to the *Gaia* EDR3 release page[5] for their values.

To the *Gaia* EDR3 data we added the probability of reliable astrometry, p, calculated by the random forest classifier, as detailed in Sect. 2, which has a range of 0–1. We include four of the values from the posterior distance PDF determined in Sect. 2.2: the median distance `dist_50`, its 1-σ upper and lower bounds (`dist_16`, `dist_84`) and the `dist_1` value, which is the 1% distance probability and used in the selection of the GCNS.

We include radial velocities included in *Gaia* EDR3 (Lindgren et al. 2021b), which are 125 354 entries; from the radial velocity experiment (Kunder et al. 2017), which contributes 2520 entries; and from a 5″cone search for each entry on the SIMBAD database[6]: 12 852 entries. From the RAVE and SIMBAD entries we removed 130 radial velocities that were >800 km s^{-1} and 4937 objects without positive uncertainties or without reference. The total number of entries with a radial velocity is 135 790 in the full sample, 82 358 of which are in the GCNS.

We also provide magnitudes from external optical, near-infrared, and mid-infrared catalogues. The optical magnitudes are GUNN g, r, z, i from, in preference order, the Panoramic Survey Telescope and Rapid Response System first release (hereafter PS1, Chambers et al. 2016), the Sloan Digital Sky Survey 13th data release (Albareti et al. 2017), and the SkyMapper Southern Survey (Wolf et al. 2018). The near-infrared magnitudes J, H, K are from the Two Micron All Sky Survey (Skrutskie et al. 2006), the mid-infrared magnitudes $W1$ and $W2$ from the CATWISE2020 release (Eisenhardt et al. 2020), and $W3$ and $W4$ from the ALLWISE data release (Cutri et al. 2013). All external matches came from the *Gaia* cross-match tables (Marrese et al. 2019), except for the CATWISE2020 catalogue as it is not included for *Gaia* EDR3. For this catalogue we used a simple nearest-neighbour cone search with a 5″limit. We emphasise that these magnitudes are provided to have a record of the value we used in this paper and to enable a simple direct use of the GCNS. If a sophisticated analysis is required that wishes to exploit the external photometry, we recommend to work directly with the external catalogues that also have quality flags that should be consulted.

For analysis of the GCNS in a galactic framework, we require coordinates (X, Y, Z) the coordinates in a barycentric rest frame positive towards the Galactic centre, positive in the direction of rotation, and positive towards the north Galactic pole, respectively. When we ignore the low correlation between the equatorial coordinates, the (X, Y, Z) and their one-sigma bounds can be calculated using the distance estimates from Sect. 2 and their Galactic coordinates.

We inferred space velocities in the Galactic reference frame U, V, W using a Bayesian formalism. Our model contains a top layer with the parameters that we aim to infer (distances and space velocities), a middle layer with their deterministic transformations into observables (parallaxes, proper motions, and radial velocities), and a bottom layer with the actual observations that are assumed to be samples from multivariate (3D) Gaussian distributions with full covariance matrices between parallaxes and proper motions, and an independent univariate Gaussian for the radial velocity. We assumed the classical deterministic relations that define space velocities in terms of the observables (*Gaia*

[4] Both tables available at the CDS.

[5] https://www.cosmos.esa.int/web/gaia/early-data-release-3

[6] Set of Identifications, Measurements and Bibliography for Astronomical Data, http://SIMBAD.u-strasbg.fr

Table 2. Content of the GCNS and rejected dataset with the first selected object as example.

Parameter	Unit	Comment	Example
source_id	...	*Gaia* EDR3 source ID	2875125810310195712
ra	deg	Right ascension (ICRS, epoch 2016.0)	0.0157909
ra_error	mas	Uncertainty	0.16
dec	deg	Declination (ICRS, epoch 2016.0)	34.1883005
dec_error	mas	Uncertainty	0.13
parallax	mas	*Gaia* EDR3 parallax	20.194
parallax_error	mas	*Gaia* EDR3 parallax uncertainty	0.225
pmra*	mas yr^{-1}	*Gaia* EDR3 Proper motion in RA	−227.366
pmra*_error	mas yr^{-1}	*Gaia* EDR3 RA proper motion uncertainty	0.206
pmdec	mas yr^{-1}	*Gaia* EDR3 Proper motion in Dec	−56.934
pmdec_error	mas yr^{-1}	*Gaia* EDR3 Dec proper motion uncertainty	0.159
phot_g_mean_mag	mag	*Gaia* G Band magnitude	8.3483
phot_g_mean_flux_over_error	mag	*Gaia* G flux to flux uncertainty ratio	6895.11
phot_bp_mean_mag	mag	*Gaia* BP Band magnitude	8.6769
phot_bp_mean_flux_over_error	mag	*Gaia* BP flux to flux uncertainty ratio	3384.69
phot_rp_mean_mag	mag	*Gaia* RP Band magnitude	7.8431
phot_rp_mean_flux_over_error	mag	*Gaia* RP flux to flux uncertainty ratio	3544.43
phot_robust_bp_rp_excess		Ratio of the sum of the BP and RP flux to the G flux	1.2100
ruwe		Renormalised unit weight error	14.26
ipd_frac_multi_peak		Fraction of windows with multiple peaks	0
adoptedRV	km s^{-1}	Adopted Radial Velocity from EDR3 or literature	−29.94
adoptedRV_error	km s^{-1}	Uncertainty in adopted RV	0.89
adoptedRV_refname		ADS Bibcode for RV	2018A&A...616A...1G
radial_velocity_is_valid		T/F Flag to indicate if RV is in eDR3	T
GCNS_prob		Probability 0 to 1 of having reliable astrometry	1.00
WD_prob		Probability 0 to 1 of being a white dwarf	1.00
dist_1	kpc	1st percentile of the distance PDF, used in GCNS selection	0.04833
dist_16	kpc	16th percentile of the distance PDF, 1σ lower bound	0.04901
dist_50	kpc	50th percentile of the distance PDF, the median distance	0.04952
dist_84	kpc	84th percentile of the distance PDF, 1σ upper bound	0.05007
xcoord_50	pc	x coordinate in the Galactic frame using dist_50, median coordinate	–15.72239
xcoord_16	pc	x coordinate 1σ lower bound	–15.55850
xcoord_84	pc	x coordinate 1σ upper bound	–15.89664
ycoord_50	pc	y coordinate in the Galactic frame using dist_50, median coordinate	41.02444
ycoord_16	pc	y coordinate 1σ lower bound	40.59680
ycoord_84	pc	y coordinate 1σ upper bound	41.47911
zcoord_50	pc	z coordinate in the Galactic frame using dist_50, median coordinate	–22.85814
zcoord_16	pc	z coordinate 1σ lower bound	–22.61987
zcoord_84	pc	z coordinate 1σ upper bound	–23.11148
uvel_50	km s^{-1}	Velocity in the Galactic frame, direction positive x	–61.07
uvel_16	km s^{-1}	Velocity 1σ lower bound	–61.69
uvel_84	km s^{-1}	Velocity 1σ upper bound	–60.43
vvel_50	km s^{-1}	Velocity in the Galactic frame, direction positive y	–5.58
vvel_16	km s^{-1}	Velocity 1σ lower bound	–6.39
vvel_84	km s^{-1}	Velocity 1σ upper bound	–4.88
wvel_50	km s^{-1}	Velocity in the Galactic frame, direction positive z	12.81
wvel_16	km s^{-1}	Velocity 1σ lower bound	12.43
wvel_84	km s^{-1}	Velocity 1σ upper bound	13.24
NAME_GUNN		Name from the PanSTARRS/SDSS/SkyMapper survey	1237663235523739680
REFNAME_GUNN		ADS Bibcode Gunn bands	2017ApJS..233...25A
gmag_GUNN	mag	GUNN G Band magnitude (SDSS:g, Skymapper: g_psf)	12.388
e_gmag_GUNN	mag	Uncertainty GUNN G Band magnitude (SDSS:err_g, Skymapper:e_g_psf)	0.007
rmag_GUNN	mag	GUNN R Band magnitude (SDSS:r, Skymapper: r_psf)	12.293
e_rmag_GUNN	mag	Uncertainty GUNN R Band magnitude (SDSS:err_r, Skymapper:e_r_psf)	0.008
imag_GUNN	mag	GUNN I Band magnitude (SDSS:i, Skymapper: i_psf)	12.445
e_imag_GUNN	mag	Uncertainty GUNN I Band magnitude (SDSS:err_i, Skymapper:e_i_psf)	0.008
zmag_GUNN	mag	GUNN Z Band magnitude (SDSS:z, Skymapper: z_psf)	9.007
e_zmag_GUNN	mag	Uncertainty GUNN Z Band magnitude (SDSS:err_z, Skymapper:e_z_psf)	0.001
NAME_2MASS		2mass name	00000410+3411189
j_m_2MASS	mag	2MASS J band magnitude	7.249
j_msig_2MASS	mag	Uncertainty 2MASS J band magnitude	0.017
h_m_2MASS	mag	2MASS H band magnitude	6.940
h_msig_2MASS	mag	Uncertainty 2MASS H band magnitude	0.016
k_m_2MASS	mag	2MASS K band magnitude	6.885
k_msig_2MASS	mag	Uncertainty 2MASS K band magnitude	0.017
NAME_WISE		WISE Name	J000003.81+341117.9
w1mpro_pm_WISE	mag	CATWISE W1 Band magnitude	7.249
w1sigmpro_pm_WISE	mag	Uncertainty CATWISE W1 Band magnitude	0.020
w2mpro_pm_WISE	mag	CATWISE W2 Band magnitude	6.922
w2sigmpro_pm_WISE	mag	Uncertainty CATWISE W2 Band magnitude	0.008
w3mpro_WISE	mag	ALLWISE W3 Band magnitude	6.883
w3sigmpro_WISE	mag	Uncertainty ALLWISE W3 Band magnitude	0.016
w4mpro_WISE	mag	ALLWISE W4 Band magnitude	6.824
w4sigmpro_WISE	mag	Uncertainty ALLWISE W4 Band magnitude	0.085

coordinates, parallaxes and proper motions, and radial velocities), which we explicitly develop in Appendix B. We neglect here for the sake of simplicity and speed the uncertainties in the celestial coordinates and their correlations with parallaxes and proper motions. The full covariance matrices are given by the catalogue uncertainties and correlations.

We used the same empirical prior for the distance as described in Sect. 2.2 and defined three independent priors for the space velocities U, V, and W (see Appendix B for details). In all three cases we use a modified Gaussian mixture model (GMM) fit to the space velocities found in a local (140 pc) simulation from the Besançon Galaxy model (Robin et al. 2003). The number of GMM components is defined by the optimal Bayesian information criterion. The modification consists of decreasing the proportion of the dominant Gaussian component in each fit by 3% and adding a new wide component of equal size centred at 0 km s^{-1} and with a standard deviation of 120 km s^{-1} to allow for potential solutions with high speeds typical of halo stars that are not sufficiently represented in the Besançon sample to justify a separate GMM component. We then used *Stan* (Carpenter et al. 2017) to produce 2000 samples from the posterior distribution and provide the median U, V, W, and their one-sigma upper and lower bounds in the output catalogue with suffixes vel_50, vel_16, and vel_84, respectively.

4. Catalogue quality assurance

4.1. Sky variation

In this section we discuss the completeness of the GCNS in the context of the full *Gaia* EDR3. In particular, we examine the changes in completeness limit with the direction on the sky as a result of our distance cut and as a result of separation of sources.

4.1.1. G magnitude limits over the sky

One of the main drivers of the completeness is the apparent brightness of a source on the sky. It can be either too bright, such that the CCDs are overexposed, or it can be too faint, such that it can hardly be picked up from background noise. For *Gaia* EDR3 the G magnitude distribution is depicted in Fig. 5 for the sources that have both a G and a parallax measurement. At the bright end, we have a limit at about 3 mag, and at the faint end, the magnitude distribution peaks at 20.41 mag (the mode), which indicates that not all sources at this magnitude are recovered by *Gaia* because otherwise the source count would still rise.

First- and second-order effects arise from the underlying source density, for example, if there are too many sources for *Gaia* to process, $\sim 10^6$ per deg^2 (de Bruijne 2012), then sources with brighter on-board G magnitude estimate are prioritised; and the scanning law, for instance, expected scans per source, vary over the sky, which can improve coverage for fainter sources. Because the latter effect is complex to simulate (Boubert & Everall 2020), we employed an empirical approach using the `gdr2_completeness`[7] `python` package (Rybizki & Drimmel 2018). We essentially focused on the G magnitude distribution per HEALpix (Górski et al. 2005), but used percentiles instead of the mode as an estimator of the limiting magnitude because the mode is noisy in low-density fields and prone to biases. Red clump stars towards the bulge or the Magellanic clouds can produce a mode in the distribution at brighter magnitudes, for example (cf. discussion in Sect. 3.2 of Rybizki et al. 2020).

[7] https://github.com/jan-rybizki/gdr2_completeness

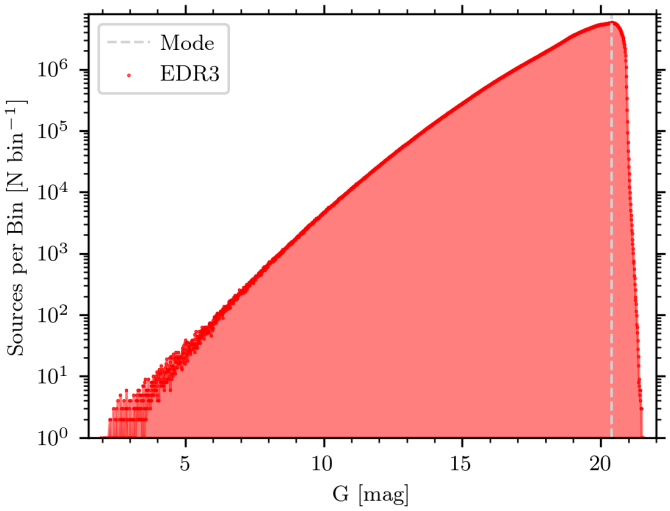

Fig. 5. G magnitude distribution for all sources in *Gaia* EDR3 that have a G magnitude and a parallax measurement (the bin size is 0.01 mag). The mode is indicated as a grey dashed line at 20.41 mag.

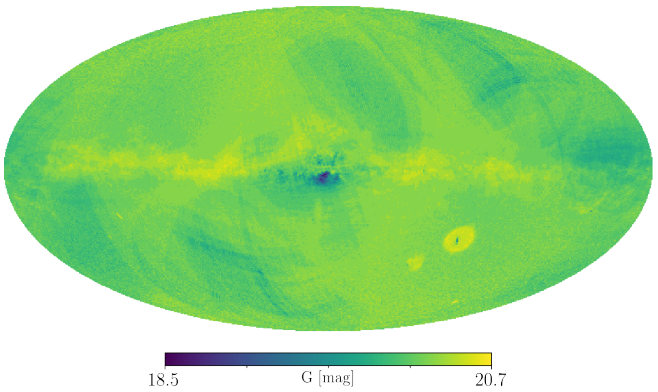

Fig. 6. 80th percentile of the G magnitude distribution per level 7 HEALpix over the sky as a Mollweide projection in Galactic coordinates. The Galactic centre is in the middle, and the longitude increases to the left.

We decided which percentile of the magnitude distribution was used. Limits of $G = 20.28$ and $G = 20.54$ encompass 80% and 90% of the sources, respectively. These limits are approximately at the left and right edge of the grey line in Fig. 5 denoting the mode at $G = 20.41$, which includes 85% of the sources. We expect a reasonable cut for most lines of sight to be between these values. We show the resulting empirical magnitude limit map in HEALpix level 7 for sources with G and parallax measurement in *Gaia* EDR3 for the 80th percentile in Fig. 6. Scanning law patterns as well as the high-density areas of the bulge and the Large Magellanic Cloud can be seen. Sources with even fainter magnitudes still enter the catalogue, but they do not represent the complete underlying population of sources at these magnitudes. These sources instead enter the *Gaia* EDR3 catalogue in a non-deterministic fashion as a consequence of the imprecise on-board G magnitude estimate. We provide the empirical G magnitude limit map including all percentiles at the HEALpix fifth level as a table available at the CDS because this is used in Sect. 5.

An external validation of our usage of percentiles as a proxy for completeness limits can be achieved by comparing *Gaia* EDR3 results cross-matched with PS1 sources classified

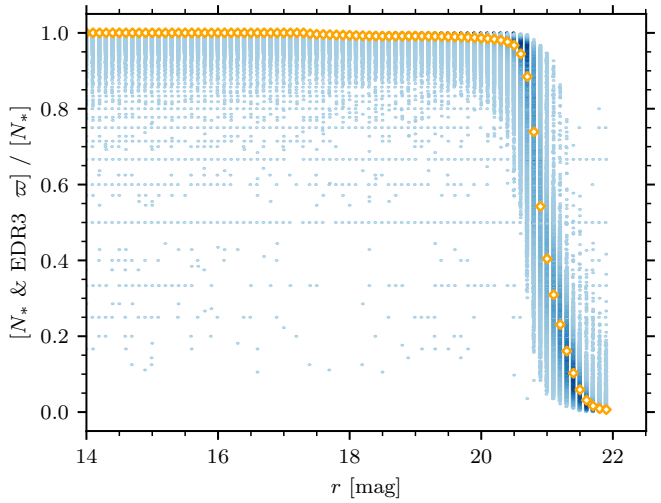

Fig. 7. Each point is the ratio of objects classified as stellar in PanStarrs with *Gaia* EDR3 parallaxes to all objects classified as stellar per level 6 HEALpixel binned in r magnitudes, as discussed in Sect. 4.1.1 The orange diamonds represent the median value for all HEALpixels.

as stars. A PS1 star is defined as an object with a probability from Tachibana & Miller (2018) greater than 0.5. We compare the full PS1 footprint and make two assumptions: that the PS1 is complete in the relevant magnitude range, and that $r \sim G$. This assumption means that the limit of the PS1 is significantly fainter than the *Gaia* limit, and for all but the reddest objects, the median $r - G$ is zero. When we assume this a simple cut at $G = (19.9, 20.2, 20.5)$ mag, which is the mean magnitude limit of the 70th, 80th, and 90th percentile map, this results in a source-count averaged completeness of 97, 95, and 91%. Using our map at 70th, 80th, and 90th percentiles, we find an 98, 97, and 95% completeness. Figure 7 shows the ratio of PS1 stellar sources with *Gaia* EDR3 parallaxes to all PS1 stellar sources in bins of magnitudes in level 6 HEALpixels. The median ratio is 99% until 19.5, drops to 95% at 20.5 (slightly different to the above G because it is averaged across the sky), and quickly sinks to 50% at 21.0.

4.1.2. Volume completeness with M_G

With regard to volume completeness per absolute magnitude, which needs to be corrected for when a luminosity function is constructed, as we do in Sects. 5.2 and 5.8.2, we take into account (a) the apparent magnitude limits and (b) the distance probability distribution. For (a) we conservatively employed the 80th percentile apparent magnitude limit map from Sect. 4.1.1 per level 5 HEALpix. All stars that are not within these limits were excluded from the analysis. For (b) we used all of the 99 PDF samples with a distance estimate $\leq$100 pc instead of a single distance estimate per source, for example by the median distance.

On the selected samples, we performed our analysis (e.g. used the respective distance and G magnitude to derive M_G and counted the sources per absolute magnitude bin), finally dividing our resulting numbers by 99 to recover the true stellar numbers. Objects that are close to the 100 pc border only contribute partially to our analysis, down-weighted by the probability mass, which resides within 100 pc. Similarly, owing to the distance PDF samples, individual sources can contribute to different absolute magnitude bins.

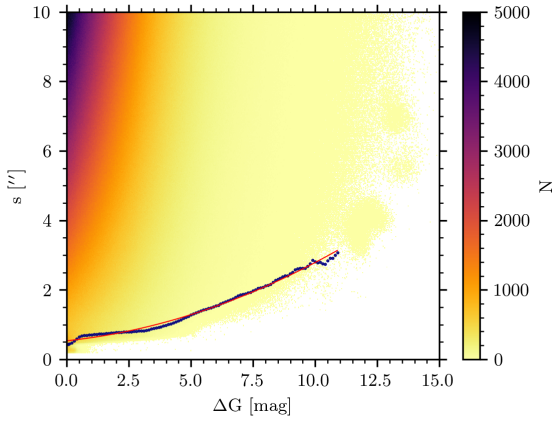

Fig. 8. Magnitude difference, ΔG, vs. angular separation, s, of all objects in *Gaia* EDR3 colour-coded by density in [0.02, 0.02] bins. The blue points represent the 99.5 percentiles of the separations, and the red line is a fit to these values and is reported in Eq. (2).

4.1.3. Contrast sensitivity

The resolving power of the *Gaia* instrument of two sources that lie close together in the sky mainly depends on the angular separation and the magnitude difference (de Bruijne et al. 2015) and is called contrast sensitivity, see Brandeker & Cataldi (2019) for a *Gaia* DR2 determination. In dense regions we especially lose faint sources due to this effect (Rybizki et al. 2020), which directly affects our ability to resolve binaries. We empirically estimated this function using the distribution of close pairs from the full *Gaia* EDR3.

In Fig. 8 we plot the angular separation of entries in the *Gaia* EDR3 as a function of the magnitude difference. The blue points are the 99.5 percentiles of the separations binned in overlapping magnitude bins of 0.2 mag in the magnitude range 0–11 mag. We adopted these percentiles as the minimum resolvable separation, $s_{\rm min}$, and therefore the dependence on the magnitude difference, ΔG, is approximated by the red line,

$$s_{\rm min} = 0.532728 + 0.075526 \cdot \Delta G + 0.014981 \cdot (\Delta G)^2. \quad (2)$$

The structure and over-densities in Fig. 8 after the 12th magnitude are due to the gating and windowing effects for bright objects observed by *Gaia*.

4.2. Comparison to previous compilations

The Set of Identifications, Measurements and Bibliography for Astronomical Data (SIMBAD) database provides information on astronomical objects of interest that have been studied in scientific articles. All objects in this database have therefore been individually vetted by a professional in some way, and while the census is not complete because not all objects have been studied, the contamination is low. From this database we retrieved all stars with a parallax larger than 8 mas through the following query performed with the TAP service[8]:

```
SELECT main_id, plx_value, plx_bibcode,
    string_agg(bibcode||';'||plx,';')
FROM basic LEFT JOIN mesPlx on oid\,{=}\,oidref
WHERE plx_value>8
GROUP BY main_id, plx_value, plx_bibcode
```

[8] https://simbad.u-strasbg.fr/simbad/sim-tap

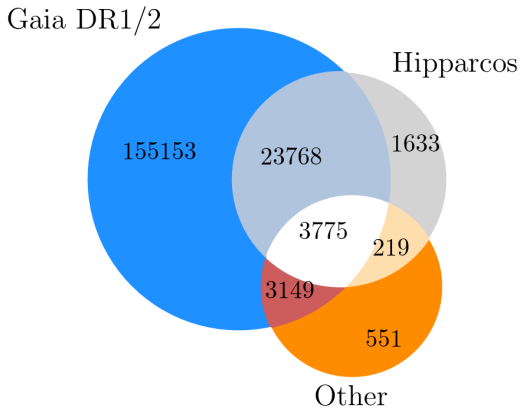

Fig. 9. Content of the 8 mas sample from the SIMBAD query. The number of stars is given in different samples depending on the origin of their parallax. The ratio of the area for the three main primary circles is proportional to the ratio of the square root of the total number of objects per sample.

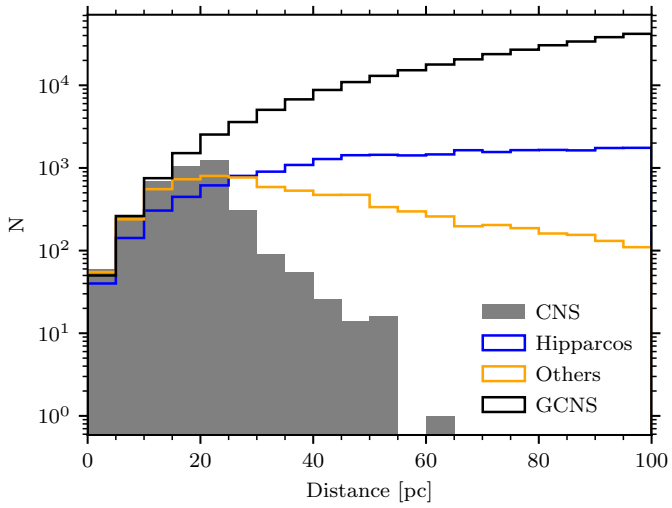

Fig. 10. Distance distributions of the GCNS compared to previous compilations. The distance is computed as the inverse of the parallax, taken from the respective catalogue. The y-axis is a log scale.

The 8 mas limit, or 125 pc, was chosen at this stage because we will further cross-match with the GCNS and expect some of the sources to have a new *Gaia* parallax above 10 mas, which means that they enter the 100 pc sample, or vice versa. This query returned 189 096 objects. Eight hundred and thirty-nine objects in binary systems are duplicates: they have one entry as a multiple system, plus one or two (or even three) entries for the individual components (e.g. α Cen is listed three times, first as a system, but then α Cen A and α Cen B are also listed individually). Moreover, obvious errors are, for example, HIP 114176, 2MASS J01365444-3509524, and 2MASS J06154370-6531528, which last case is a galaxy.

This leaves a sample of 188 248 objects. Most of them, ~98%, have parallaxes from *Gaia* DR1 (566 objects; Gaia Collaboration 2016) and *Gaia* DR2 (184 584 objects; Gaia Collaboration 2018b), and ~2% have parallaxes from HIPPARCOS (2534 objects; Perryman et al. 1997; van Leeuwen 2007). The few remaining objects (564) are from other trigonometric parallax programs (e.g. van Altena et al. 1995; Smart et al. 2013; Dittmann et al. 2014; Martin et al. 2018).

SIMBAD does not systematically replace `plx_value` by the most recent determination, but prefers the value with the lowest measurement uncertainty. In particular, for 693 very bright stars from this query, the astrometric solution of HIPPARCOS is chosen over that of *Gaia* DR2.

Our SIMBAD query also gives all existing trigonometric parallax measurements (from the table `mesPlx`) for each star. Figure 9 shows the content of the SIMBAD query in terms of the number of stars and the origin of their parallax. It shows that the SIMBAD 8 mas sample has mostly been fed by *Gaia*. For this reason, we first compared GCNS with the compilation, excluding the objects for which only a *Gaia* parallax is available (blue sample in Fig. 9).

Next we compared GCNS with the full *Gaia* DR2 data (and not only with the stars listed in SIMBAD, which are about half of the full *Gaia* DR2 catalogue). Within the 100 pc sphere, the total number of objects having an astrometric parallax determination consequently increases from 26 536 stars prior to *Gaia* to 300 526 stars in GCNS with `dist_50`< 0.1, or 301 797 stars when each source with `dist_1`< 0.1 is counted and weighted by its probability mass (see Sect. 4.1.2).

Figure 10 shows the distribution in distance of the GCNS catalogue, of the HIPPARCOS catalogue, and of all objects having a parallax from other programmes (mainly ground-based), prior to *Gaia*. The Catalogue of Nearby Stars (Gliese & Jahreiß 1991) is also shown. Although CNS is based on ground-based programs, CNS contains more stars in some bins than are listed in SIMBAD. The main reason is that CNS lists all the components in multiple systems, whereas SIMBAD has only one entry for the systems with one parallax measurement.

In what follows, we compare *Gaia* EDR3 objects with a parallax $\hat{\varpi} > 8$ mas and a probability $p > 0.38$ (see Sect. 2) with previous compilations. We first cross-matched *Gaia* EDR3 with the SIMBAD sample (excluding exoplanets and stars with only a *Gaia* parallax), and we retrieved 94% of the objects. The *Gaia* EDR3 adds 402 stars to the 100 pc sphere and removes 318 stars. Some stars have very different parallax determinations. For instance, HD 215415 has a parallax of 79.78 ± 21.65 mas from HIPPARCOS and 10.32 ± 0.05 mas from *Gaia* EDR3. This is a double star, which may question the validity of the measurements.

SIMBAD contains 1 245 objects with `plx-value` >10 mas that are not in *Gaia* EDR3. They are shown in Fig. 11. Some of them are too faint or bright or are binaries, but for some there is no clear consistent reason why they are missing. In particular, half of them are in *Gaia* DR2. We provide a table[9] of these missing objects, in which we also included 4 stars within 10 pc and 9 confirmed ultra-cool dwarfs with a parallax measurement from *Gaia* DR2 that were not individually in SIMBAD, but were confirmed independently.

We next compared our sample with *Gaia* DR2, to which the same process of training set construction and random forest classifier creation was applied for quality assurance (maintaining the same overall choices, but adapting the feature space to those available in *Gaia* DR2). We selected the stars with parallax $\hat{\varpi} > 8$ mas and a probability $p > 0.43$, which is the optimum threshold given by the ROC for *Gaia* DR2. With this, we retrieved 95% of the *Gaia* DR2 stars in *Gaia* EDR3. Figure 12 shows the comparison in the parallax distributions. It is clear that *Gaia* DR2 has significantly more false entries and spurious large parallaxes. One reason that a GCNS was not attempted with *Gaia* DR2 was that the amount of false objects in the original data was excessive; the selection procedure would have found

[9] Available at the CDS.

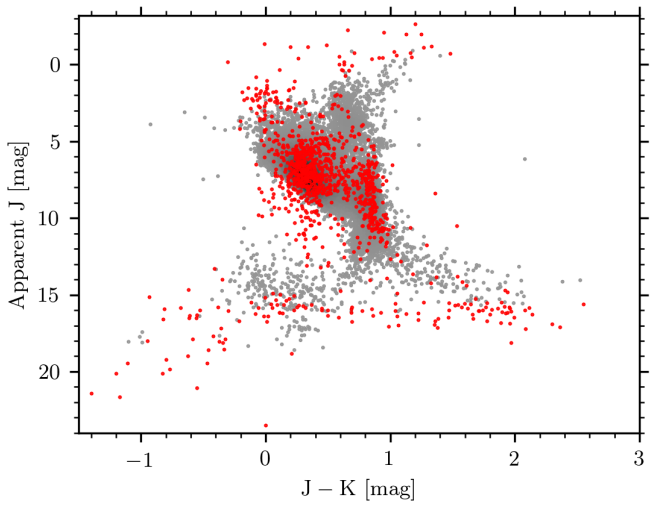

Fig. 11. J vs. $J-K$ of the SIMBAD 100 pc sample before *Gaia*. Grey dots: stars found in *Gaia* EDR3. Red dots: stars not found in *Gaia* EDR3.

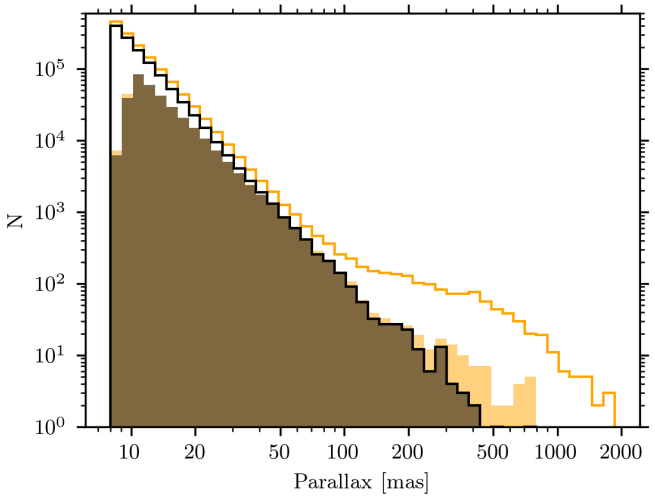

Fig. 12. Parallax distribution in *Gaia* DR2 with $\hat{\varpi} > 8$ mas (empty, orange), *Gaia* DR2 with `dist_1`< 0.1 and $p > 0.43$ (filled, orange), *Gaia* EDR3 with $\hat{\varpi} > 8$ mas (empty, black), and GCNS (filled, black).

15 objects with $\hat{\varpi} > 500$ mas if we had made a GCNS with *Gaia* DR2.

Within 100 pc (i.e. `dist_50` <0.1 kpc), 7079 stars published in *Gaia* DR2 are not found in *Gaia* EDR3, and 8760 stars in *Gaia* EDR3 are not in *Gaia* DR2. Their position in the CAMD, in a G versus distance diagram, and on the sky in Galactic coordinates is shown in Fig. 13. The left panels show the stars in *Gaia* DR2 that are not *Gaia* EDR3. Some of them are very faint ($M_G > 21$), very close ($\hat{\varpi} > 100$ mas), and as already noted, they are false entries in *Gaia* DR2. Stars in the left part of the main sequence are also suspicious because it appears that they are also located along scanning law patterns, as revealed by the lower left panel of Fig. 13. In particular, the clump around $M_G = 8$ corresponds to the over-density of stars around $G = 12$ in the G versus distance diagram (see the middle left panel). We suspect that the pile-up and gap at $G \sim 13$ are related to the effects in *Gaia* DR2 of changing window class across this magnitude range, see Evans et al. (2018), Riello et al. (2018), and Carrasco et al. (2016). Even given these known artefacts in the *Gaia* datasets, we are still left with ~3400 *Gaia* DR2 stars that are located along the main sequence for which there is no evident reason why they are not in *Gaia* EDR3.

In contrast, the G versus distance diagram is smoother for the stars that are found in *Gaia* EDR3 but not in *Gaia* DR2 (middle right panel). The faint stars (at $G > 20$) correspond to the WDs and low-mass stars. Thousands of new candidates are thus expected for these faint objects (see Sects. 5.8 and 5.4). The CAMD in the top right panel is coloured as a function of the parameter `ipd_frac_multi_peak`. It provides the fraction of windows as percentage from 0 to 100 for which the algorithm has identified a double peak, that is, a high value indicating that the object is probably a resolved double star. Many stars lie at the right side of the main sequence, where we expect overluminous binary systems to lie, and a significant fraction have a consistently high probability to be binaries.

Many objects with low `ipd_frac_multi_peak` values remain as outliers with red colours, probably due to inconsistent photometry (see also Sect. 5.5). The sky map (lower right panel) shows regions (in particular, $l \simeq 240°$ and $b \simeq 45°$) in which stars are found in *Gaia* EDR3 , but not *Gaia* DR2.

4.3. 10 pc sample

As an illustration, we detail the 10 pc sample. The SIMBAD query returns 393 objects (excluding exoplanets) From this list we removed 14 duplicates, one error (HIP 114176), and *Gaia* DR2 4733794485572154752, which we suspect to be an artefact that lies in front of a globular cluster. We added the multiple brown dwarf Luhman 16 AB at 2 pc (Luhman 2013) and 2MASS J19284155+2356016, a T6 at 6 pc (Kirkpatrick et al. 2019). The resulting 10 pc sample contains 378 objects, 307 of which are in the GCNS. The new *Gaia* EDR3 parallax places LP 388-55 outside the 10 pc sphere, and HD 260655 enters this sphere.

The GCNS lists the first individual parallax measurements for five stars in systems within the 10 pc sample: HD 32450B, CD-37 10765B, the WD o 2 Eri B, Wolf 424 B, and one star in the μ Her system, separated by 0.6″from μ.02 Her. This means that 312 stars are located within 10 pc in the GCNS.

We removed all giants, WDs, and peculiar or uncertain types from the full set of SIMBAD spectral types, and we find a calibration between the median absolute magnitude, M_G, and each spectral class. With this calibration for the SIMBAD entries with spectral types, we predicted their apparent G magnitudes. Ten of the objects missing in GCNS are stars that are too bright (Sirius, Fomalhaut, α Cen A and B, Vega, Procyon, Altair, Mizar, χ Dra, and HD 156384), 33 objects are T and Y brown dwarfs and are too faint, as are probably 2 late-L dwarfs. However, of the remaining 26 objects, 15 have *Gaia* DR2 parallaxes. We note that 21 of these 26 objects are either spectroscopic binaries or in close binary systems that will give high residuals with a single-star solution, and for this reason, they may not have passed the *Gaia* five-parameter solution quality assurance tests. Five objects remain (HD 152751, G 24-16, IRAS 21500+5903, SCR J1546-5534, and BPS CS 22879-0089) for which we do not have an obvious reason to explain the lack of a *Gaia* EDR3 five-parameter solution.

The resulting 10 pc sample contains 383 objects with a parallax determination: 376 stars from SIMBAD minus LP 388-55, plus Luhman 16 AB, 2MASS J19284155+2356016, HD 260655, and the five companion stars with a first parallax determination from *Gaia* EDR3 . There are also known unresolved binary systems (Procyon, η Cas, ξ UMa, etc.), and as there will undoubtedly be new systems (e.g. see Vrijmoet et al. 2020),

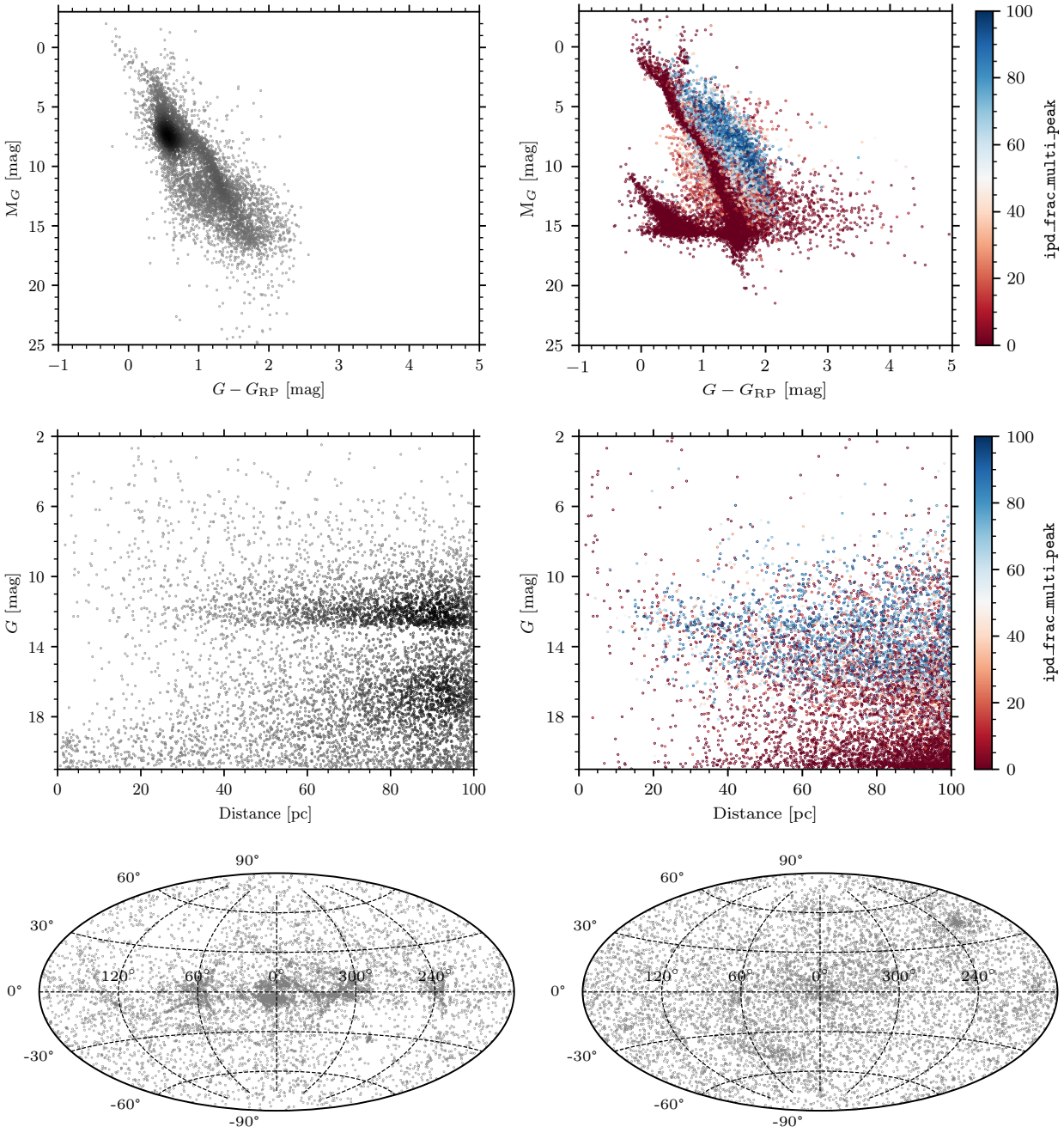

Fig. 13. Comparison between *Gaia* DR2 and *Gaia* EDR3, *from top to bottom*, in the CAMD, in a G vs. distance diagram, and on the sky in galactic coordinates. *Left*: stars in *Gaia* DR2 not found in *Gaia* EDR3. *Right*: stars in *Gaia* EDR3 not in *Gaia* DR2. *Upper and middle right panels*: coloured with the `ipd_frac_multi_peak`. This parameter, available in *Gaia* EDR3, provides the fraction of windows as percentage from 0 to 100 for which the detection algorithm has identified a "double peak", meaning that it was probably a visually resolved double star (either just a visual double or a real binary).

we counted unresolved systems as one entry. The T/Y types will not be complete in this list, for instance, the 16 T6 to Y2 brown dwarfs that are not included in this list have a parallax larger than 100 mas from Kirkpatrick et al. (2019), and more ultra-cool Y-dwarfs are expected to be discovered. These 383 objects can be retrieved by selecting entries from the GCNS with `parallax` >= 100 mas (312 objects) and that in the file with missing objects that we provide have `plx_value` >= 100 mas (71 objects). To provide a starting point for estimating the number of objects expected within 100 pc in the next section, we estimate that the number of objects with $M_G < 15.5$ within 10 pc is 316.

4.4. Consistency check with the 10 pc sample

In order to check for the plausibility of the total number of sources that are classified as good by the random forest described in Sect. 2, we used the Einasto law with the maximum a posteriori values of the parameters inferred in Sect. 5.1 to produce synthetic samples of sources with uniform densities in planes parallel to the Galactic plane. We produced an arbitrary number of samples and then scaled the numbers to match the observed number of sources within 10 pc. As the *Gaia* 10 pc sample will be missing bright sources and to avoid a possible circular reasoning, we used a census of known sources inside 10 pc with an absolute

magnitude brighter than 15.5 mag regardless of whether the sources are detected by *Gaia*.

In our simulations we assumed for the sake of simplicity that the binary population properties are dominated by the M spectral type regime. We set a binarity fraction of 25% and a distribution of binary separations (a) that is Gaussian in logarithmic scale, with the mean and standard deviation equal to 1: $\log_{10}(a) \sim \mathcal{N}(\mu = 1, \sigma = 1)$ (see Robin et al. 2012; Arenou 2010, and references therein). We assumed that the orientation of the orbital planes are random and uniform in space, giving rise to the usual law for the inclinations i with respect to the line of sight given by $p(i) \sim \sin(i)$. Furthermore, we assigned a magnitude difference between the two components (we did not include higher order systems) based on the relative frequencies encountered in the GCNS and discussed in Sect. 5.6. Based on the separations and inclinations, we computed the fraction f of the orbit where the apparent angular separation of the binary components is larger than the angular separation in Eq. (2). The probability of detecting the binary system as two separate sources was then approximated with the binomial distribution for a number of trials equal to 22 (which is the mode of the distribution of the number of astrometric transits in our dataset) and success probability f. This is an optimistic estimate because it assumes that one single separate detection suffices to resolve the binary system.

Using the procedure described above, we generated ten simulations with 40 million sources each, distributed in a cube of $110 \times 110 \times 110$ pc. From each simulation we extracted the number of sources within 10 and 100 pc (N_{10} and N_{100}, respectively) and the ratio between the two (N_{100}/N_{10}). The average value of this ratio from our simulations is 878.2 ± 28.2. When we apply this scale to the observed number of sources within 10 pc (316 sources), the expected number of sources in the GCNS selection is $277\,511 \pm 8911$. This prediction has to be compared with the number of sources in the GCNS catalogue with an absolute magnitude brighter than 15.5 and within 100 pc. In order to obtain this number, we proceeded as described in Sect. 4.1.2 and obtained a total number of sources of 282 652, which agrees well with the prediction given the relatively large uncertainties and the fact that the number of sources within 10 pc (316) is itself a sample from a Poisson distribution. It has to be borne in mind, however, that the expected number (277 511) does not take incompleteness due to variations across the sky of the G magnitude level or due to the contrast sensitivity into account. GeDR3mock simulations show that 1.8k sources are fainter than the *Gaia* EDR3 85th percentile magnitude limits (cf. Sect. 4.1.1). Additional 0.3k sources are lost due to the contrast sensitivity, which will be a lower limit because GeDR3mock does not include binaries, so that this is only the contribution of chance alignments in crowded regions.

4.5. Contamination and completeness

As described in Lindegren et al. (2018), every solution in *Gaia* is the result of iteratively solving with different versions of the input data and varying the calibration models. The final solutions do not use all the observations and not all solutions are published, many quality assurance tests are applied to publish only high-confidence solutions. Internal parameter tests that were applied to publish the five-parameter solution in *Gaia* EDR3 were $G <= 21.0$; `astrometric_sigma5d_max` $< 1.2 \times 10^{0.2\max(6-G,0,G-18)}$ mas; `visibility_periods_used` > 8; `longestsemiMajorAxis` of the position uncertainty ellipse<= 100 mas; and `duplicateSourceID` $= 0$. The tests were calibrated to provide a balance between including poor solutions and rejecting good solutions for the majority of objects, that is, distant, slow-moving objects whose characteristics are different from those of the nearby sample. In the current pipeline, the astrometric solution considers targets as single stars, and for nearby unresolved or close binary systems the residuals of the observed motion to the predicted motion can be quite large, so that this causes some nearby objects to fail the `astrometric_sigma5d_max` test.

For example, as we saw in Sect. 4.2, we expect 383 objects within the 10 pc sample. When the 35 L/T objects that we consider too faint are removed, 348 objects remain that *Gaia* should see (we include the bright objects for the purpose of this exercise). Twenty-six of these 348 objects do not have five-parameter solutions in *Gaia* EDR3 because they fail the solution quality checks. The fact that many of the lost objects were in spectroscopic or close binary systems is also an indication that the use of a single-star solution biases the solutions for the nearby sample. If we take these numbers directly, this loss is still relatively small: 26 of 348, or 7.4%. While this loss is biased towards binary systems, it probably does not depend on direction and the loss will diminish as the distance increases because the effect of binary motion on the solutions decreases. The excess of objects found in the GCNS compared to the prediction in Sect. 4.4 supports this conclusion, and the comparison of objects found in SIMBAD to those in the GCNS shows that only 6% are missing, therefore we consider the 7.4% as a worst-case estimate of the GCNS stellar incompleteness.

Section 4.1.1 showed that the mode, or peak, of the apparent G distribution is at $G = 20.41$ mag, which includes 85% of the sources. The median absolute magnitude of an M9 is $M_G = 15.48$ mag, which would translate into $G = 20.48$ mag at 100 pc; therefore we should see at least 50% of the M9-type stars at our catalogue limit. Our comparison to the PS1 catalogue indicates that *Gaia* EDR3 is 98% complete at this magnitude. As discussed further in Sect. 5.4, the complete volume for later spectral types becomes progressively smaller, but for spectral types up to M8, they are volume limited and not magnitude limited.

We lose small numbers of objects because we started with a sample that was selected with $\varpi > 8$ mas, which from the GDR3Mock is estimated to be 55. We will lose objects that are separated by less than 0.6″ due to contrast sensitivity (Sect. 4.1.3), which for chance alignments from the GDR3Mock we have estimated to be 300 sources, but it will be much higher for close binary systems and will bias our sample to not include these objects. Finally, we lose objects that are incorrectly removed because they have $p < 0.38$. Based on Table 1 and Sect. 2, we estimate this to be approximately 0.1% of the good objects.

The incompleteness for non-binary objects to spectral type M8 is therefore dominated by the 7.4% of objects for which *Gaia* does not provide a parallax. For objects later than M8, the complete volume decreases, as shown in Sect. 5.4. We did not consider unresolved binary systems, which are considered in Sects. 5.7 and 5.6.

We also considered the contamination of the GCNS. There are two types of contamination: objects that pass our probability cut but have poor astrometric solutions, and objects that are beyond our 100 pc limit. The contamination of the good solutions is evident in the blue points that populate the horizontal feature at $M_G = 15$–16 mag and between the main and WD sequence (Fig. 2, right panel). These are faint objects ($G > 20.$ mag) that lie at the limit of our distance selection (`dist_50` $= 80$–120 pc),

for example with a distance modulus of ∼5 mag), and that therefore populate the M_G >15 mag region. These faint objects have the lowest signal-to-noise ratio, and their parameters, used in the random forest procedure, therefore have the largest uncertainties. Because objects with poor astrometric solutions were accepted, we estimate this contamination based on Table 1 to be ∼0.1%, the false positives. This means about 3000 objects for the GCNS.

The contamination by objects beyond the 100 pc sphere can be estimated by summing the number of distance probability quantiles inside and outside 100 pc. We find that 91.2% of the probability mass lies within 100 pc and the rest outside. This means 29k sources, or 9%. The use of the full distance PDF will allow addressing this possible source of bias in any analysis.

These known shortcomings should be considered when the GCNS is used. If the science case requires a clean 100 pc sample, where no contamination is a priority and completeness is of secondary importance, objects with a `dist_50` < 0.1 kpc should be selected from the GCNS. If the science case requires a complete sample, all objects with `dist_1` < 0.1 kpc should be selected and then weighted by the distance PDF. When a clean photometric sample is required, the photometric flags should be applied, which we did not exploit to produce this catalogue. In the next section we investigate a number of science questions, for which we apply different selection procedures to the catalogue and use the distance PDF in different ways to illustrate some optimal uses of the GNCS.

5. GCNS exploitation

5.1. Vertical stratification

In this section we study the vertical stratification as inferred from the GCNS volume-limited sample. We did this using a relatively simple Bayesian hierarchical model that we describe in the following paragraphs. First we describe the data we used to infer the vertical stratification parameters, however. The data consist of the latitudes, observed parallaxes, and associated uncertainties of the sources in the GCNS with observed parallaxes greater than 10 mas. In order to include the effect of the truncation in the observed parallax, we also used the number of sources with observed parallaxes between 8 and 10 mas and their latitudes (but not their parallaxes). The reasons for this (and the approximations underlying this choice) will become clear after the inference model specification. The assumptions underlying the model listed below.

1. The data used for inference represent a sample of sources with true parallaxes larger than 8 mas. This is only an approximation, and we know that the observed sample is incomplete and contaminated. It is incomplete for several reasons, but in the context of this model, the reason is that sources with true parallaxes greater than 8 mas may have observed parallaxes smaller than this limit due to observational uncertainties. It is also contaminated because the opposite is also true: true parallaxes smaller than 8 mas may be scattered in as a result of observational uncertainties as well. Because this effect is stronger than the first reason and more sources lie at larger distances, we expect fewer true sources with true parallaxes greater than 8 mas (at distances closer than 125 pc) than were found in the GCNS.
2. The source distribution in planes parallel to the Galactic plane is isotropic. That is, the values of the true Galactic Cartesian coordinates x and y are distributed uniformly in any such plane.
3. The measurement uncertainties associated with the observed Galactic latitude values are sufficiently small that their effect on the distance inference is negligible. Uncertainties in the measurement of the Galactic latitude have an effect on the inference of distances because we expect different distance probability distributions for different Galactic latitudes. For example, for observing directions in the plane that contains the Sun, the true distance distribution is only dictated by the increase in the volume of rings at increasing true distances (all rings are at the same height above the Galactic plane and therefore have the same volume density of sources), while in other directions the effect of increasing or decreasing volume densities due to the stratification modifies the true distance distribution.
4. Galactic latitudes are angles measured with respect to a plane that contains the Sun. This plane is parallel to the Galactic plane but offset with respect to it by an unknown amount.
5. Parallax measurements of different sources are independent. This is known to be untrue but the covariances amongst *Gaia* measurements are not available and their effect is assumed to cancel out over the entire celestial sphere.

For a constant volume density ρ and solid angle $\mathrm{d}\Omega$ along a given line of sight, the probability density for the distance r is proportional to r^2. In a scenario with vertical stratification, however, the volume density is not constant along the line of sight but depends on r through z, the Cartesian Galactic coordinate. For the case of the Einasto stratification law (Einasto 1979) that is used in the Besançon Galaxy model (Robin et al. 2003), the distribution of sources around the Sun is determined by the ϵ parameter (the axis ratio) and the vertical offset of the Sun , $Z_\odot$, with respect to the fundamental plane that defines the highest density. The analytical expression of the Einasto law for ages older than 0.15 Gyr is

$$\rho \propto \rho_0 \cdot \exp\left(-\left(0.5^2 + \frac{a^2}{R_+^2}\right)^{\frac{1}{2}}\right) - \exp\left(-\left(0.5^2 + \frac{a^2}{R_-^2}\right)^{\frac{1}{2}}\right), \tag{3}$$

where $a^2 = R^2 + \frac{z^2}{\epsilon^2}$, R is the solar galactocentric distance, z is the Cartesian Galactic coordinate (which depends on the Galactic latitude b and the offset as $z = r \cdot \sin(b) + Z_\odot$), ϵ is the axis ratio, and we used the same values as in the Besançon model, $R_+ = 2530$ pc and $R_- = 1320$ pc. The value of ϵ in general depends on age. We assumed a single value for all GCNS sources independent of the age or the physical parameters of the source such as mass, effective temperatures, and evolutionary state.

In our inference model we have the vertical stratification law parameters (ϵ and $Z_\odot$) at the top. We defined a prior for the ϵ parameter given by a Gaussian distribution centred at 0.05 and with a standard deviation equal to 0.1, and a Gaussian prior centred at 0 and with a standard deviation of 10 pc for the offset of the Sun with respect to the Galactic plane. Then, for a given source with Galactic latitude b, the probability density for the true distance r is given by

$$p(r \mid \epsilon, Z_\odot) \propto \rho(z(r) \mid \epsilon, Z_\odot) \cdot r^2. \tag{4}$$

Equation (4) is the natural extension of the constant volume density distribution of the distances. Finally, for N observations of the parallax $\hat{\varpi}_i$ with associated uncertainties σ_{ϖ_i}, the likelihood is defined as

$$\mathcal{L} = \prod_1^N p(\hat{\varpi}_i \mid r_i, \epsilon, Z_\odot) = \prod_1^N \mathcal{N}(\hat{\varpi}_i \mid r_i, \sigma_{\varpi_i}), \tag{5}$$

where $\mathcal{N}(\cdot \mid \mu, \sigma)$ represents the Gaussian (or normal) distribution centred at μ and with standard deviation σ, and we have introduced the assumption that all parallax measurement are independent. The model is defined by the stratification parameter ϵ, the (also) global parameter $Z_\odot$, and the N true distances to individual sources r_i. With this, the posterior distribution for the full forward model can be expressed as

$$p(\epsilon, Z_\odot, \boldsymbol{r} \mid \hat{\boldsymbol{\varpi}}) \propto \prod_1^N p(\hat{\varpi}_i \mid r_i, \epsilon, Z_\odot) \cdot p(r_i \mid \epsilon, Z_\odot) \cdot p(\epsilon) \cdot p(Z_\odot), \tag{6}$$

where bold symbols represent vectors. For the sake of computational efficiency, we marginalised over the N individual distance parameters r_i and inferred only the two global parameters ϵ and $Z_\odot$,

$$\begin{aligned} p(\epsilon, Z_\odot \mid \hat{\boldsymbol{\varpi}}) &\propto \int p(\epsilon, Z_\odot, \boldsymbol{r} \mid \hat{\boldsymbol{\varpi}}) \cdot \mathrm{d}\boldsymbol{r} \\ &= \prod_1^N \int_0^{r_{\max}} p(\hat{\varpi}_i \mid r_i, \epsilon, Z_\odot) \cdot p(r_i \mid \epsilon, Z_\odot) \cdot \mathrm{d}r_i \cdot p(\epsilon) \cdot p(Z_\odot), \end{aligned} \tag{7}$$

where $r_{\max}$ represents the assumed maximum true distance in the sample of sources that defines the dataset.

The model described so far relies on the assumption that the dataset used for the evaluation of the likelihood is a complete and uncontaminated set of the sources with true distances between 0 and $r_{\max}$. The selection of this dataset from the observations is impossible, however. On the one hand, the posterior distances derived in Sect. 2.2 assume an isotropic prior, and a selection based on it would therefore be (mildly) inconsistent. The inconsistency is minor because the directional dependence of the prior is a second-order effect with respect to the dominant r^2 factor. It is also problematic because a source with a posterior median slightly greater than 100 pc would be left out of the sample even though it has a relatively high probability to be inside, and vice versa for sources with posterior medians slightly smaller than 100 pc. On the other hand, a selection based on the observed parallax (e.g. defined by $\hat{\varpi} > 10$ mas) is different from the sample assumed by the model (which is defined by all true distances being within the 100 pc boundary). We decided to modify the model to account for a truncation in the space of observations for illustration purposes. It exemplifies an imperfect yet reasonable way to deal with such truncations.

We inferred the model parameters from the set of sources with observed parallaxes $\hat{\varpi} > 10$ mas, but we modified the likelihood term in order to include the truncation of observed parallaxes. The dataset upon which our model infers the stratification parameters was defined by all sources classified as good astrometric solutions with the random forest described in Sect. 2, for example with $p > = 0.38$. We assumed that the total number of sources with true distances smaller than 125 pc is the same as that with observed parallaxes greater than or equal to 8 mas. This is an approximation because we know that in general, the true number will be smaller due to the effect of the measurement uncertainties scattering more external sources in than internal sources out. However, the true number cannot be estimated without knowing the stratification parameters. It is possible to infer the total number as another model parameter, but that is beyond the scope of this demonstration paper. We modified the likelihood term to include the effect of the truncation as follows.

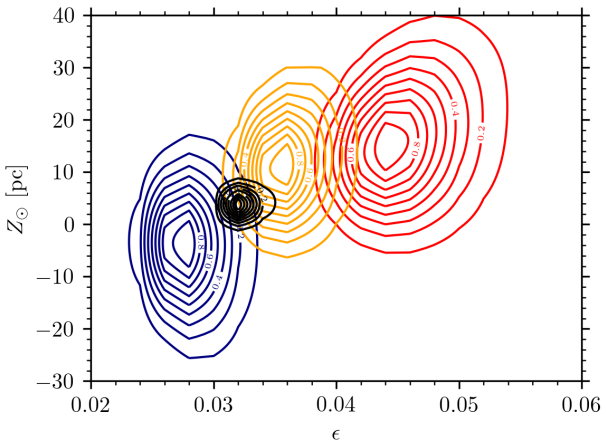

Fig. 14. Posterior probability density for the Einasto law ϵ parameter and the solar z coordinate for the entire GCNS (black) and for three segments along the main-sequence distributions from left to right: early spectral types before the turn-off point (blue), spectral types G and early K (orange), and M-type stars (red).

The likelihood term was divided into two contributions that distinguish sources with $\hat{\varpi} > 10$ mas and sources with smaller parallaxes ($8 < \hat{\varpi} < 10$ mas). For the former, the likelihood term was exactly as described by Eq. (5). For the latter we only retained their Galactic latitudes, the fact that their observed parallaxes are smaller than 10 mas, and the total number of sources, but not the parallax measurements themselves. The new likelihood term is

$$\begin{aligned} \mathcal{L} &= p(\hat{\boldsymbol{\varpi}} \mid r, \epsilon, Z_\odot) \\ &= \prod_1^{N_{\mathrm{obs}}} p(\hat{\varpi}_i \mid r_i, \epsilon, Z_\odot) \cdot \prod_1^{N_{\mathrm{miss}}} \int_{-\infty}^{10} p(\hat{\varpi}_i \mid r_i, \epsilon, Z_\odot) \cdot \mathrm{d}\hat{\varpi}_i, \end{aligned} \tag{8}$$

where N_{obs} is the number of sources with observed parallaxes $\hat{\varpi} \geq 10$ mas, and N_{miss} is the number of sources with observed parallaxes in the range from 8 to 10 mas. With this new likelihood expression, we can proceed to calculate posterior probability densities for a given choice of priors.

Figure 14 shows the posterior density contours for the ϵ parameter and the solar coordinate $z = Z_\odot$ under the Einasto model described above for the entire sample, small contours in the middle, and for three separate subsamples along the main sequence. The maximum a posteriori values of the model parameters for the full GCNS sample are $\epsilon = 0.032$ and $Z_\odot = 4$ pc. Figure 14 shows that the hot population (defined as the main-sequence segment brighter than $M_G = 4$) seems characterised by a smaller ϵ parameter (with a maximum a posteriori value of 0.028) and a vertical coordinate $Z_\odot = -3.5$, whereas the middle ($4 < M_G < 7$) and cool ($12 < M_G < 15$) segments of the main sequence are characterised by higher values of ϵ (0.036 and 0.044, respectively) and $Z_\odot$ coordinates larger than the inferred value for the full sample (11.5 and 15, respectively). For comparison, the values used in the Besançon Galaxy model (Robin et al. 2003) range between 0.0268 for stars younger than 1 Gyr and 0.0791 for those with ages between 7 and 10 Gyr. The parameters inferred in this section are fully consistent with these values given that the data samples are not characterised by a single age but contain sources with a continuum of ages (younger on average for the hot segment, and increasingly older for the middle or cool segments) determined by the local star formation history and kinematical mixing.

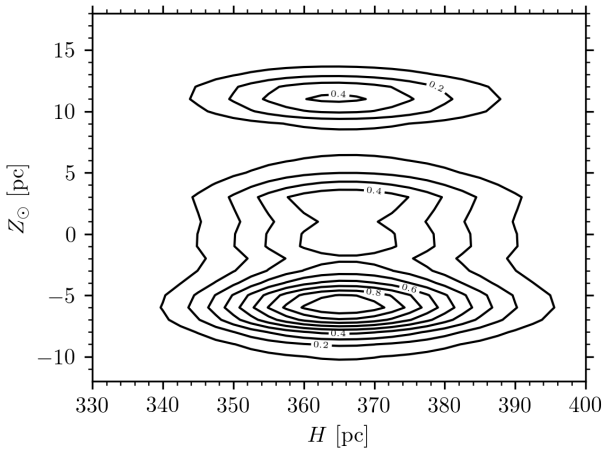

Fig. 15. Posterior probability density for the scale height H and solar z coordinate with respect to to the Galactic plane inferred from the hierarchical Bayesian model.

The values discussed in the previous paragraph did not take into account that the objects used to infer the stratification parameters include sources that do not belong to the thin disc. In order to assess the effect of the presence of thick disc stars in our dataset on the inferred parameter values, we applied the same method to an augmented dataset with an additional 6.6% of sources (see Sect. 5.3.2 for a justification of this value) distributed uniformly in the three Galactic Cartesian coordinates. This was an upper limit to the effect because thick-disc stars are also vertically stratified. It results in a small shift of the inferred parameters characterised by a maximum a posteriori value of ϵ and of the vertical coordinate $Z_\odot$ of 0.034 and 3.5 pc, respectively.

We also applied the formalism described above to the alternative stratification model defined by the exponential decay with scale height H (see Dobbie & Warren 2020, for a recent application of Bayesian techniques to a set of analytical stratification laws that includes the exponential model). The prior probability for the scale height is defined as an exponential distribution with scale 1000 pc, and that for the offset is defined as a Gaussian distribution centred at 0 and with a standard deviation of 10 pc (the same as in the case of the Einasto law). Figure 15 shows the un-normalised posterior for the model described above and parameters H and $Z_\odot$. The maximum a posteriori value of the vertical scale height is 365 pc, above the value of 300 pc commonly accepted in the literature (see Rix & Bovy 2013, and references therein), and certainly greater than more recent estimates such as those of Dobbie & Warren (2020). We interpret this difference as due to the discontinuity of the derivative of the exponential distribution at $Z = 0$ and the limited range of distances of the sample used here. If the true density distribution is smooth at that point (i.e. if the likelihood term that includes the exponential decay is not a good model of the data), then it is to be expected that the inference model favours values of H that are higher than would be inferred over larger volumes. The value of $Z_\odot$ is less constrained by the data and the marginal distribution is multi-modal, with a maximum at −6 and several local minima at positive coordinates. Given the sharp peak of the exponential distribution, we interpret the various maxima as the result of local over-densities. The negative maximum a posteriori value of $Z_\odot$ is surprising because the values found in the literature range between 5 and 60 pc, with most of the recent measurements concentrated between 5 and 30 pc (see Table 3 of Karim & Mamajek 2017, and references therein). A direct comparison of the values is difficult, however, because each measurement defines the Galactic plane in a different way. In our case, we measured the vertical position of the Sun with respect to the z coordinate of the local (within 100 pc of the Sun) maximum volume density. This does not need to coincide with the Galactic plane defined by the distribution of star counts of different stellar populations (e.g. Cepheids, Wolf-Rayet stars, or OB-type), the distribution of clusters, or the distribution of molecular gas, especially if these distributions are not local but averaged over much larger fractions of the Galactic disc.

5.2. Luminosity function

The GCNS is an exquisite dataset from which to derive the local luminosity function. This is possible for the first time using volume-limited samples with parallaxes not derived from photometric measurements that are affected by related biases (Eddington or Malmquist), and homogeneously throughout the HR diagram, from bright stars down to white dwarfs and the substellar regime. In this section, we present the luminosity function of main-sequence and giant stars.

We first removed all objects with a probability higher than 50% to be a WD as defined in Sect. 5.8. The giant branch is well separated from the main sequence in the M_G versus $G_{\rm BP} - G_{\rm RP}$ diagram. Our giant star selection follows the two conditions $M_G < 3.85$ and $G_{\rm BP} - G_{\rm RP} > 0.91$ and gives 1573 stars, which is a significant sample even given the small volume. The remaining stars are considered to belong to the main sequence. At this stage, we did not attempt to correct the luminosity function for binarity effects. We thus defined a subsample of the main sequence keeping only stars with `ipd_frac_multi_peak` = 0 corresponding to 81% of the main-sequence stars. As already mentioned, this parameter reflects the probably of being a visually resolved binary star. This filter decreases the binarity contribution and at the same time removes some of the outliers with $G - G_{\rm RP}$ colour excess whose photometry is suspected to be incorrect (see Sect. 5.5).

We determined the luminosity function using the generalised form of the $V_{\rm max}$ classical technique (Schmidt 1968). We computed the maximum volume probed at a given absolute magnitude, and corrected it to take the decrease in stellar density with increasing distance above the Galactic plane (Felten 1976; Tinney et al. 1993) into account,

$$V_{\rm max} = \Omega \frac{H^3}{\sin^3 |b|} [2 - (\xi^2 + 2\xi + 2) \exp(-\xi)], \tag{9}$$

with

$$\xi = \frac{d_{\rm max} \sin |b|}{H}, \tag{10}$$

where H is the thin-disc scale height and $d_{\rm max}$ is the maximum distance of the detection. b and Ω are the Galactic latitude and the area of the HEALpix to which the star belongs. We assumed $H = 365$ pc as derived in Sect. 5.1.

Usually, $d_{\rm max}$ was estimated for each object. Thus the object was counted as the inverse of the maximum volume $V_{\rm max}$ in which it is observed. The luminosity function is the sum over all objects within an absolute magnitude bin,

$$\Phi(M) = \sum \frac{1}{V_{\rm max}}. \tag{11}$$

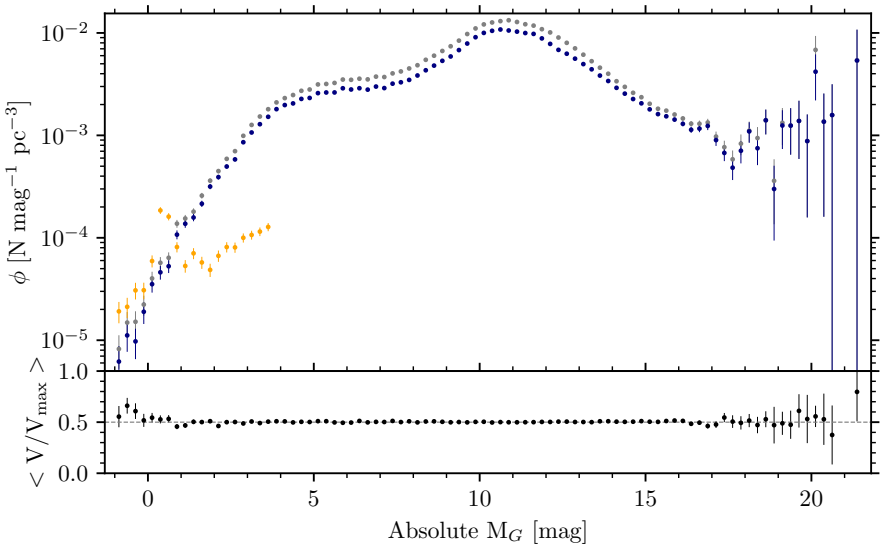

Fig. 16. *Upper panel*: luminosity function of the GCNS, with a 0.25 bin, in log scale. The upper full curve plotted in grey shows main-sequence stars. The lower full curve plotted in blue points represents main-sequence stars with `ipd_frac_multi_peak` = 0, that is, probably single stars. The small lower partially orange curve shows giants stars. The confidence intervals reflect the Poisson uncertainties. *Lower panel*: $\langle V/V_{\rm max}\rangle$ vs. M_G. The expectation value for the statistic is 0.5 for a uniform sample within the survey volume.

We followed this scheme, but as explained in Sect. 4.1.2, instead of using a single distance for each star, we considered its whole distance probability distribution, adding a contribution of 1/99th of the sum of those probabilities within 100 pc. The use of the Bayesian framework allowed us to avoid the Lutz-Kelker bias for a volume-limited sample (Lutz & Kelker 1973).

Finally, we took the 80th percentile G magnitude sky distribution at HEALpix level 5 (corresponding to an angular resolution of three square degrees per HEALpix) to apply sensitivity cuts and reach the highest completeness limit depending on the sky position. These limits have a minimum, mean, and maximum G limit of 18.7, 20.2, and 20.7 respectively.

We also computed the mean value of $V/V_{\rm max}$ in the bins of absolute magnitude, where V is the (generalised) volume in which each object is discovered, that is, the volume within the distance d to each object. This statistic $\langle V/V_{\rm max}\rangle$ should approach 0.5 for a uniformly distributed sample with equal counts in each volume. As shown in the lower panel of Fig. 16, this is the case of our sample from $M_G = 2$ to ~20.5.

The luminosity function is shown in the upper panel of Fig. 16. The luminosity function of the giant sample is shown in red. The red clump is clearly visible by the peak at $M_G = 0.4$ with $\Phi = 1.9 \pm 0.1 \times 10^{-4}$ stars pc^{-3} mag^{-1}. However, this is underestimated here because objects brighter than $G \simeq 3$ are not included in *Gaia*. The local luminosity function of fainter giants, on the red giant branch, is reliable, however. We compared our result with those obtained by Just et al. (2015) using a sample of 2660 giants from HIPPARCOS and the CNS up to 200 pc. They found a value of $\Phi = 8.3 \times 10^{-5}$ stars pc^{-3} mag^{-1} at $M_K = 1$, which roughly corresponds to $M_G = 3$, where we find a consistent $11.0 \pm 1.1 \times 10^{-5}$ stars pc^{-3} mag^{-1}.

The luminosity function of the main-sequence sample illustrates the very high precision offered by the unprecedented quality of the GCNS. The confidence intervals reflecting Poisson uncertainties are very small even at the low-mass end down to $M_G \simeq 18$ mag, corresponding to L3-L4 based on the spectral type versus M_G relation derived for SIMBAD entries as described in Sect. 4.2. The overall density is 0.081 ± 0.003 stars pc^{-3}.

This can be compared with the previous efforts made to determine the luminosity function within 25 pc based on HIPPARCOS CNS (e.g. Just et al. 2015), and ground-based observations (e.g. Reid et al. 2002, and references therein). By using a combination of HIPPARCOS and astrometric and spectroscopic observations, Reid et al. (2002) were able to derive the solar neighbourhood (25 pc) luminosity function from bright to low-mass stars, including the contribution from companions. There is an overall agreement with our determination, in particular within their confidence intervals, that can be 20 times larger than in the GCNS luminosity function (see their Fig. 8). One main difference is the double-peaked shape in their luminosity function, with one maximum at $M_V = 12.5$ mag (corresponding to our maximum at $M_G = 10.5$ mag) and a higher one at $M_V = 15.5$ mag that should stand at $M_G = 14$–14.5 mag and does not appear in the GCNS luminosity function. This second peak is poorly defined: it has a large confidence interval, for instance, and does not appear in the 8 pc luminosity function determined by Reid et al. (2003).

The high precision of the luminosity function enables searching for signatures of structures in the CAMD, such as the Jao gap. Using *Gaia* DR2, Jao et al. (2018) discovered this narrow gap (~0.05 mag) in the lower main sequence, which is hypothesised to be the result of a dip in the luminosity function associated with complex evolutionary features of stars with mass ~0.35 $M_\odot$ (MacDonald & Gizis 2018; Baraffe & Chabrier 2018). Therefore, we first inspected the lower main sequence of the complete GCNS catalogue to verify the presence of this feature, and find that the gap stands out distinctly, as depicted in the left panel of Fig. 17. By breaking down the GCNS sample into $G_{\rm BP} - G_{\rm RP}$ colour and magnitude bins according to Jao et al. (2018), their Table 1, we also confirm the largest decrement of counts around $M_G = 10.14$, or $M_{\rm RP} = 9.04$. The effects of this gap are reflected in the luminosity function, as shown in the right panel of Fig. 17 by the red line. The main sequence also shows an inflection close to the gap that is very likely the effect noted by Clemens et al. (1998). Other structure is apparent in the luminosity function that may be connected to the main-sequence structure found in Jao & Feiden (2020) as well as the more classical variations (Wielen et al. 1983; Kroupa et al. 1990), but this is beyond the scope of this contribution.

Recent works have been made to derive the luminosity function at the stellar to substellar boundary (Bardalez Gagliuffi et al. 2019) and for L to Y brown dwarfs (Kirkpatrick et al. 2019). Although the statistical noise increases in the brown dwarf regime, the luminosity function can be derived down to $M_G = 20.5$ (translating into ~L9 spectral type). Several features

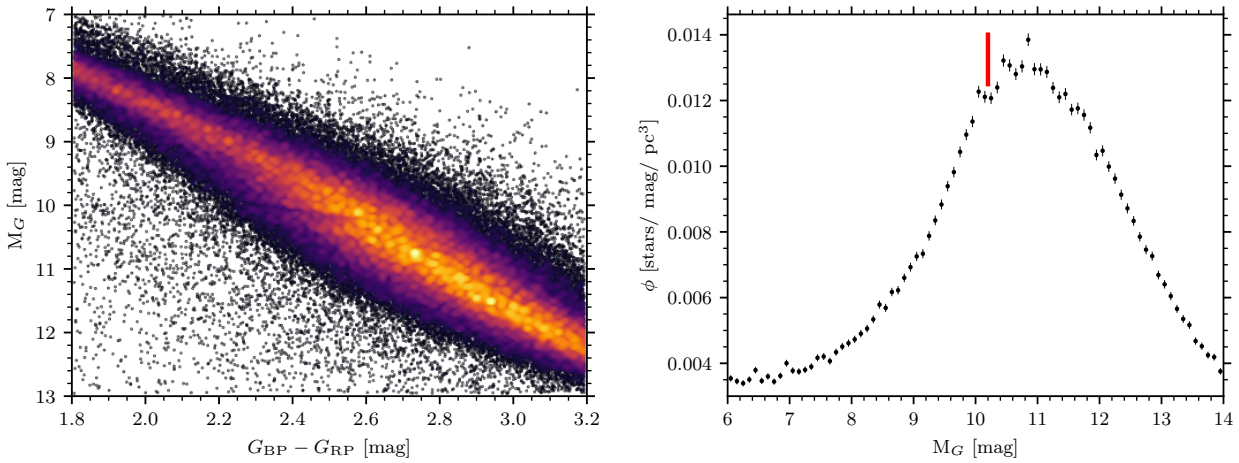

Fig. 17. *Left panel*: so-called Jao gap in the M_G vs. $G_{RP} - G_{BP}$ diagram. *Right panel*: zoom of the luminosity function from Fig. 16 computed in 0.1 magnitude bins, in linear scale. The red line indicates the position of the Jao gap.

can still be seen: a dip at $M_G = 16.4$ mag or L0; a dip at $M_G = 18.9$ mag or L5; a peak at $M_G = 20.1$ mag or L8, but with a large confidence interval. A better investigation of the possible contamination by red objects with potentially inaccurate photometry, in particular in the late-L regime, should be made before any conclusions are drawn as to the reality of these features.

However, the clear dip at $M_G = 17.6$ in Fig. 16, which is also seen at $M_{RP} = 16.11$ in the M_{RP} luminosity function, corresponding to the L3 spectral type, probably is a real feature. This can be seen in Fig. 26 from Bardalez Gagliuffi et al. (2019), for instance, where a plateau appears at $M_J \simeq 13$ to 14 (corresponding to $M_G \simeq 16$ to 18, M9 to L4), followed by an increase in luminosity function. This absolute magnitude region lies at the edge of other studies (Cruz et al. 2007; Bardalez Gagliuffi et al. 2019 for M7 to L5 and Reylé et al. 2010 for L5 to T0) using different samples. In contrast, the GCNS offers a homogeneous sample that gives confidence to the physical significance of that dip. This minimum probably is a signature of the stellar to substellar boundary because brown dwarfs are rapidly cooling down with time. Models predict that they pass through several spectral types within 1 Gyr (see e.g. Baraffe et al. 2015). Thus they depopulate earlier spectral types to go to later ones. The homogeneous dataset offered by GCNS will allow us to refine the locus of this boundary.

5.3. Kinematics

We explored the kinematics of the GCNS catalogue by restricting the sample to 74 281 stars with a valid radial velocity in *Gaia* EDR3. We used the `vel_50` Cartesian velocities (U, V, W) as determined in Sect. 3.

5.3.1. Structures in the (U, V) plane

The sample in the (U, V) plane shows several substructures, as already pointed out in early studies by Eggen, who identified numerous groups or superclusters (Eggen 1958, 1971), then from HIPPARCOS data in Skuljan et al. (1999) and Chereul et al. (1997). See Antoja et al. (2010) for a detailed historical review. These substructures were confirmed in *Gaia* DR2 (Gaia Collaboration 2018d).

In this local sample, the (U, V) plane also appears to be highly structured, as shown in Fig. 18, where we show the approximate structuralisation of the velocity space by straight

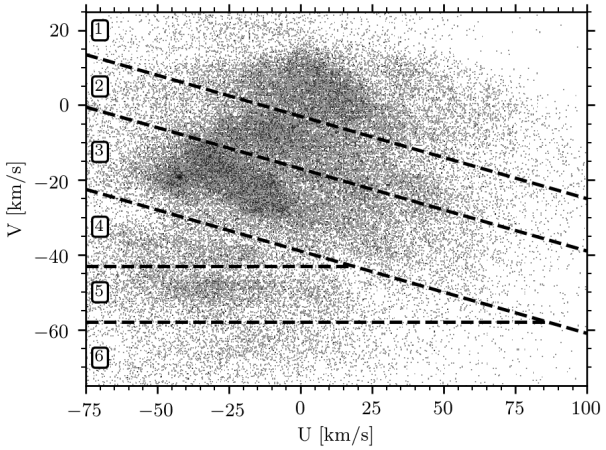

Fig. 18. GCNS stars in the (U, V) plane. We identify substructures, labelled strip 1 to 6, from top to bottom, separated by indicative straight lines: $V = 0.22 * U - 3$, $V = 0.22 * U - 17$, $V = 0.22 * U - 39$, $V = -43$, and $V = -58$.

line divisions. The three top strips have been largely studied from HIPPARCOS data, particularly by Skuljan et al. (1999) from wavelet transform, who labelled them as (1) the Sirius branch at the top (where the Sirius supercluster identified by Eggen is located), (2) the middle branch, which is less populated, and (3) the Pleiades branch, which is most populated, where the Hyades and the Pleiades groups are located. A significant gap lies just below this strip 3; it has been presented in previous studies and is nicely visible in *Gaia* DR2 data. The strips below the gap are nearly parallel to the U axis. Strips 4 and 5, seen at $V \approx -35\,\mathrm{km\,s^{-1}}$ and $\approx -45\,\mathrm{km\,s^{-1}}$, are most probably associated with the Hercules stream that was identified by Eggen (1958), where the high-velocity star ζ Hercules is located. The Hercules stream was identified at $V \approx -50\,\mathrm{km\,s^{-1}}$ by Skuljan et al. (1999), but appears itself to be substructured (Dehnen 1998) when the sample is sufficiently populated and has accurate velocities. The Gaia Collaboration (2018d) suggested that the strip at $V \approx -70\,\mathrm{km\,s^{-1}}$ (strip 6) might also be linked to the Hercules stream.

The strips are not related to cluster disruptions, as pointed out by different studies (e.g. Dehnen 1998; Antoja et al. 2010), they cover wide age ranges (Famaey et al. 2008). Some studies argued

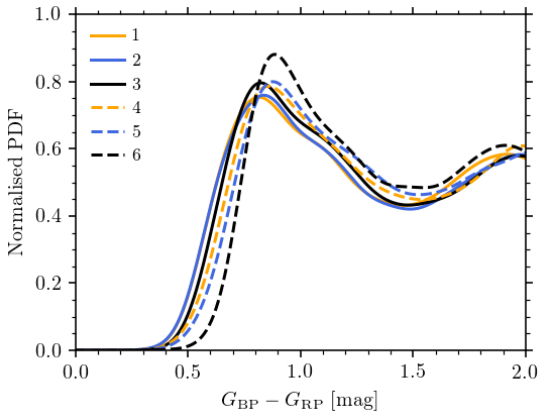

Fig. 19. KDE distribution of $G_{BP} - G_{RP}$ colour for different strips of the (U, V) plane of Fig. 18. Turn-off colours are similar for the three strips above the gap (strips 1 to 3), but older and older in strips below it (strips 4 to 6).

that they are due to resonances from either the bar, the spiral arms, or both. As a verification test, we investigated whether there is evidence of an age difference between different strips by examining their turn-off colour.

We show in Fig. 19 the KDE distribution of colours for the different strips. There is a clear colour shift of the turn-off of the different strips for those that are below the main gap (strips 4, 5 and 6 in Fig. 18), indicating that they are increasingly older when the asymmetric drift is stronger and extends farther, as expected from secular evolution. However, this global trend is superimposed on a structure that is probably due to resonances that several studies attributed to the outer Lindblad resonance of the bar (Antoja et al. 2012, 2014; Monari et al. 2017) , while others linked it to the spiral structure (Hunt et al. 2018; Michtchenko et al. 2018). There is no indication of an age dependence in our study for the three strips above the gap. They all appear to have the same turn-off colour. Therefore the structure of the velocities can be due solely to dynamical effects, such as resonances of the bar and/or the spiral (Antoja et al. 2009).

5.3.2. Stellar populations and orbits

The left panel in Fig. 20 shows the Toomre diagram of the sample. The circles with 100 and 200 km s^{-1} radii delineate thin-disc, thick-disc, and halo stars. Using these limits, we estimate that 95% of the stars belongs to the thin disc, 6.6% to the thick disc, and 0.4% to the halo. However, we show in what follows that the GCNS also contains tens of stars that visit from the central part of the Milky Way.

As discussed before, the disc kinematics is not smooth, even in the 100 pc sphere. The Toomre diagram shows many structures, but not only in the disc. The nearby halo is clumpy as well.

The diamond, with the highest retrograde velocity, shows the twin pair HD 134439 and HD 134440. They are known to be chemically anomalous stars. Their chemical compositions are close to those observed in dwarf galaxies such as Draco and Fornax, indicating an extragalactic origin (Reggiani & Meléndez 2018), and are consistent with the kinematics study of Carney et al. (1996).

The orbital parameters were computed using the online tool Gravpot16[10]. The Galactic potential we used is a non-axisymmetric potential including the bar, developed by Fernandez-Trincado (2017) and used in Gaia Collaboration (2018c) to derive orbital parameters of globular clusters and dwarf galaxies from *Gaia* DR2 data. We assumed a bar mass of 10^{10} $M_\odot$, with a pattern speed of 43 km s^{-1} kpc^{-1} and a bar angle of 20°.

The orbits integrated forward over 1 Gyr are shown in Fig. 21 in the (X, Y, Z) referential system of the Galaxy. The most numerous disc stars populate the circular orbits in the Galactic plane ($Z = 0$). Halo stars have higher eccentricities and inclinations. The central part of the (X, Y) plane is populated by the orbits of stars coming from the central regions of the Galaxy.

The 12 circled dots in Fig. 20 are the stars that we identified as related to *Gaia*–Enceladus (they are out of the panel of Fig. 18, but their velocities are centred on $U = 267 \pm 10$ km s^{-1} and $V = -221 \pm 11$ km s^{-1}.) We selected them based on their orbital parameters, in particular, in the total energy versus angular momentum (Fig. 20, right panel), where they are concentrated at $E = -156\,000$ km^{2}s^{-2} with a dispersion of 2660 km^{2}s^{-2} and $L_z = -273$ kpc km s^{-1} with a dispersion of 94 kpc km s^{-1}. These values can be compared with the last large merger event experienced by the Milky Way discovered by Helmi et al. (2018), who identified the so-called *Gaia*-Enceladus substructure with a selection of -1500 kpc km s^{-1} $< L_z <$ 150 kpc km s^{-1} and $E > -180\,000$ km^2 s^{-2}. In our local sample, the number of stars that can be attributed to Enceladus is much smaller than in the discovery paper. However, thanks to the exquisite accuracy of the parallaxes and proper motions, the structure is concentrated in orbital elements as well as velocity space. The pericentres are close to the Galactic centre (<2 kpc) and apocentres between 16 and 20 kpc (see Fig. 21, blue orbit).

The solar neighbourhood is also visited by stars from the central region of the Galaxy. The apocentres of about 40 stars in our sample lie close to the Sun, and the pericentre distance of the stars is smaller than 1 kpc (see Fig. 21, magenta orbit). They have high eccentricities, a minimum energy $E \approx -200\,000$ km^2 s^{-2}, and small angular momentum $|L_z| < 400$ kpc km s^{-1}), some are slightly retrograde.

5.3.3. Solar motion

The GCNS is well suited for measuring the solar motion relative to the local standard of rest (LSR). $U_\odot$ and $W_\odot$ velocities are easy to measure, while $V_\odot$ is subject to controversies, with values varying from 1 to 21 km s^{-1}(e.g. Dehnen & Binney 1998; Schönrich et al. 2010; Bovy et al. 2012; Robin et al. 2017). The traditional way of computing the V solar motion is to extrapolate the distribution of V as a function of U^2 to $U = 0$ for a given sample. It is expected that at $U = 0$ the mean V corresponds to the solar velocity because as a function of age the rotation of stars experiences a lag induced by the secular evolution (stars become more eccentric and the motion is less circular). The youngest stars have a lowest U velocity also because of secular evolution. Depending on the mean age of the sample, on its location (close to or farther away from the solar neighbourhood), the literature values of $V_\odot$ have been disputed. In this sense, the younger the stars are in the sample, the better the sample is for measuring $V_\odot$. However, the youngest stars are also those that experience clumping because their kinematics are far from relaxed, so they do not represent the LSR well in a general manner. Therefore we considered the whole GCNS sample as representative of the LSR, and compared the median

[10] https://gravpot.utinam.cnrs.fr/

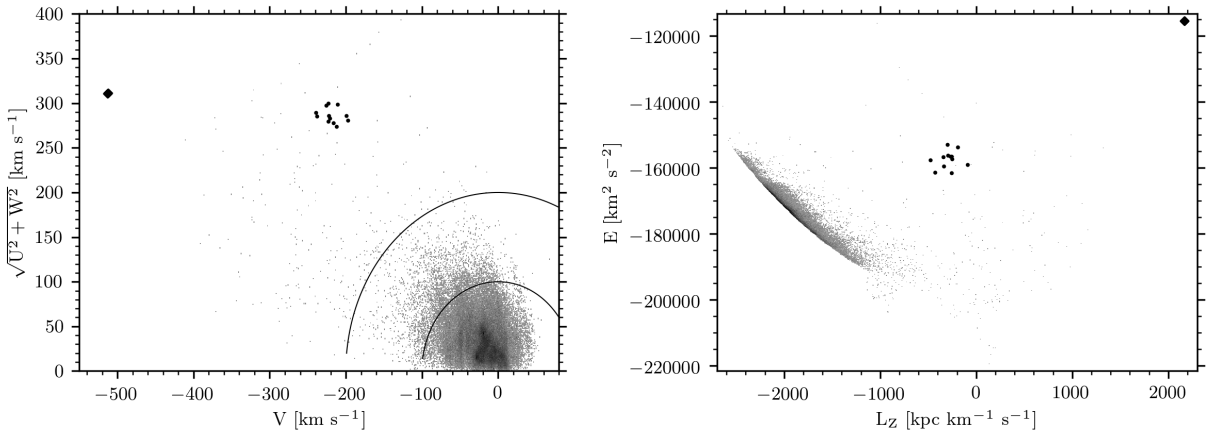

Fig. 20. *Left panel*: Toomre diagram for all the GCNS entries. The diamond symbols are the binary HD 134439/HD 134440, the circles are the *Gaia*-Enceladus group members. *Right panel*: energy vs. angular momentum for the GCNS. The symbols are the same same as the left panel.

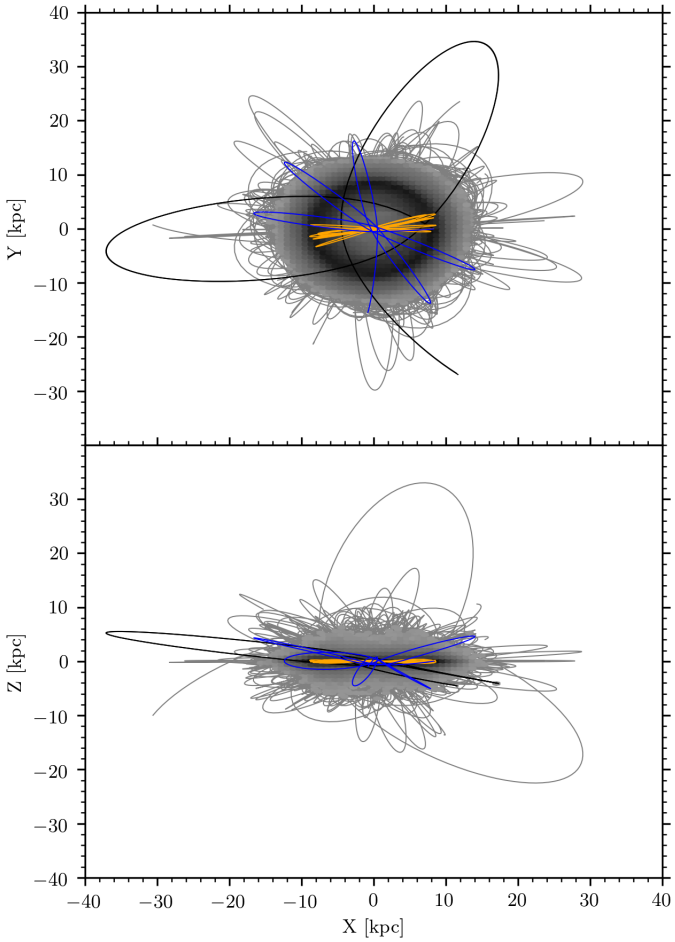

Fig. 21. Orbits of the GCNS sample. The orbits are computed over 1 Gyr and plotted in the referential system of the Galaxy. Orbits are highlighted for a few stars: the halo pair HD 134439 and HD 134440 (black), a star from the Enceladus group (blue), and a star with minimum energy (orange) coming from the central regions of the Galaxy.

V velocity of the sample with a simulation of the Besançon Galaxy model. The simulation is the same as we used for unresolved stellar multiplicity (see Sect. 5.7). The kinematics was computed from a self-consistent dynamical solution using an approximate Staeckel potential (Bienaymé et al. 2015, 2018), while the kinematic parameters of the disc populations (mainly the age-velocity dispersion relation) were fitted to *Gaia* DR1 and RAVE data as described in Robin et al. (2017). We used of the latest determination of the rotation curve from Eilers et al. (2019) based on *Gaia* DR2 . In contrast to the *Gaia* object generator (GOG; Luri et al. 2014), we added observational noise from a simplified model of the parallax and photometric uncertainties to these simulations using the equations given on the ESA website[11] as a function of magnitude and position on the sky.

To study the solar motion, we considered the GCNS sample, selecting only stars having $G < 13$ as a clear cut for stars with good radial velocities that can then be applied simply to the simulation. We first plot histograms of the distribution in U, V, and W for the sample in Fig. 22. We over-plot the simulation where the solar velocities are assumed to be (11.3, 6 ,7) km s^{-1}. While the U and W velocities are well represented by the simulation, this is not the case for the V velocity distribution, which shows significant non-Gaussianity. The clusters were not removed from the observed sample. For comparison we also over-plot a simulation assuming alternative $V_\odot$ of 12 km s^{-1} (Schönrich et al. 2010).

The non-Gaussianity of the V distribution was already known and is partly due to secular evolution and asymmetric drift, as expected even in an axisymmetric galaxy, and partly due to substructures in the (U, V) plane that are probably associated with resonances due to the bar and the spiral arms. It is beyond the scope of this paper to analyse and interpret the detailed features. However, we explored the V velocity distribution slightly more and plot the median V as a function of $G_{BP} - G_{RP}$. The blue stars are expected to be younger in the mean, while redder stars cover all disc ages. Figure 23 shows that at $G_{BP} - G_{RP} > 0.8$, the median V velocity is constant at about $V = -20$ km s^{-1}, while at $G_{BP} - G_{RP} < 0.7$ there is a shift of the median V. We over-plot a simulation with $V_\odot = 6$ and 12 km s^{-1}. The data agree well with the simulation when a solar $V_\odot$ of 6 km s^{-1} is assumed, especially for $G_{BP} - G_{RP} < 1.5$ mag, which dominate the sample. For redder stars the mean $V_\odot$ varies more with colour, it is therefore less secure to define the mean solar motion in the region. However, the median V velocity even for a local sample is a complex mix of substructures with different mean motions and of the expected asymmetric drift. With these solar velocities, the solar apex is towards $l = 31.5°$, $b = 27.2°$.

[11] https://www.cosmos.esa.int/web/gaia/science-performance

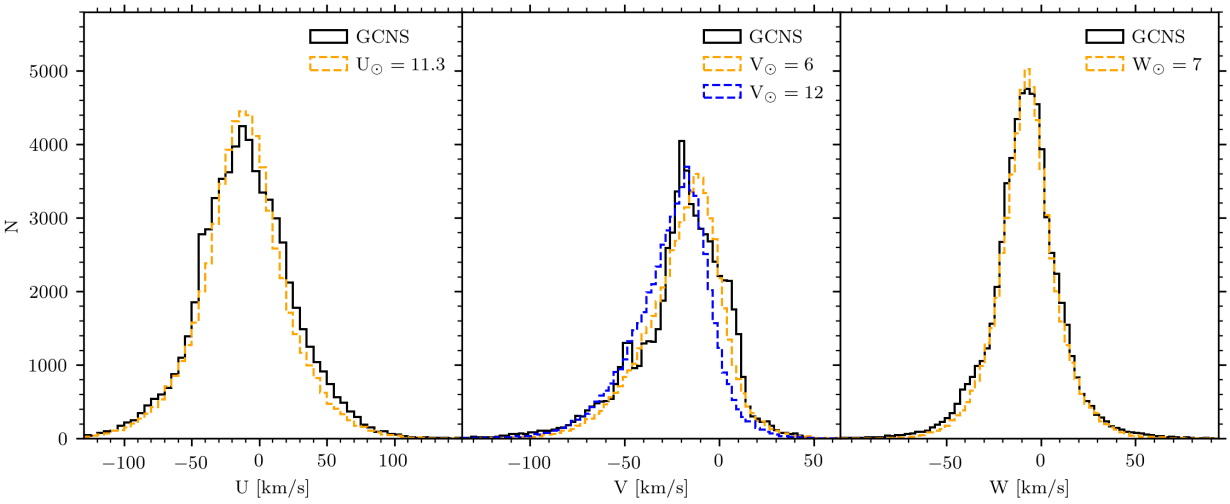

Fig. 22. Histograms of U, V, and W velocity in GCNS catalogue with $G < 13$ mag (black lines) compared with simulations with distance <100 pc and $G < 13$ (dashed lines). The simulations assume solar velocities: $U = 11.3$, $V = 6$ km s^{-1} (dashed orange line), and 12 km s^{-1} (dashed blue line), $W = 7$ km s^{-1}. The cluster members have not been removed from the data.

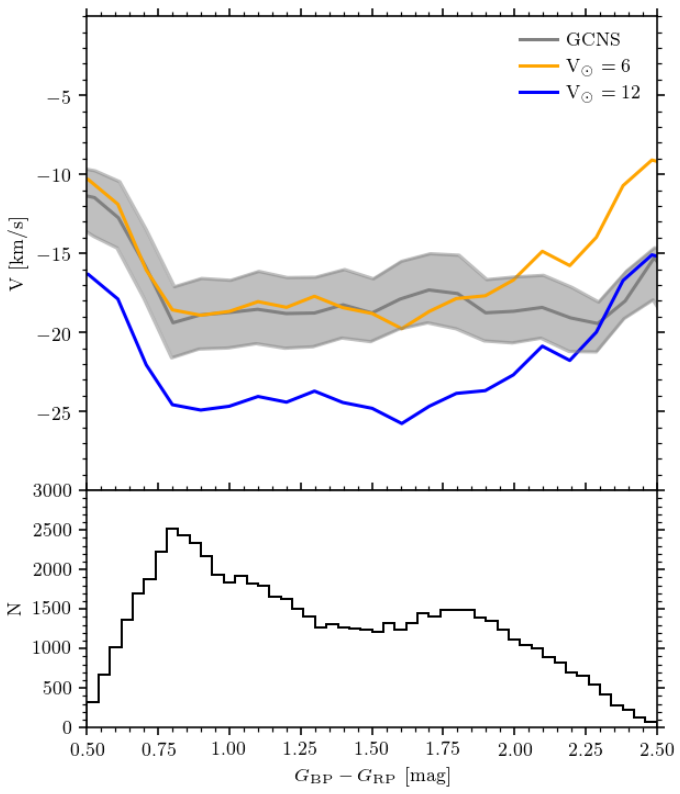

Fig. 23. *Upper panel*: median velocity as a function of $G_{BP} - G_{RP}$ colour for the GCNS sample with $G < 13$. The data with quantiles 0.45 and 0.55 are plotted in grey and the median is shown in black. The simulation was made with $V_\odot$ of 6 km s^{-1} (orange) and 12 km s^{-1} (blue). *Lower panel*: histogram of the colour distribution in *Gaia* EDR3.

We also considered the vertex deviation concept defined as the apex of the velocity ellipsoid when it is slightly rotated and does not point towards the Galactic centre, as would be expected in an axisymmetric disc. It has long been seen that young populations experience a vertex deviation, at least locally, which has been interpreted as an effect of the spiral perturbation, for instance, some theoretical analysis can be found in Mayor (1970) and Creze & Mennessier (1973). The distribution in the (U, V) plane clearly shows that strip 1 shows an inclination of the ellipsoid in the mean. However, this might also be due to the substructures in the (U, V) plane rather than a true deviation of the vertex because this strip appears to be made of the superposition of at least three superclusters or groups.

5.4. Stellar to substellar boundary

The nearby sample is particularly important for the ultra-cool dwarfs (UCDs), which are the lowest-mass, coldest, and faintest products of star formation, making them difficult to detect at large distances. They were defined by Kirkpatrick et al. (1997) as objects with spectral types M7 and later, through L, T, and Y types, have masses $M < 0.1 M_\odot$, and effective temperatures <2700 K. UCDs are of particular interest because they include both very low-mass stars that slowly fuse hydrogen, and brown dwarfs, which have insufficient mass (below about 0.075 $M_\odot$) to sustain hydrogen fusion in their cores, and slowly cool down with time.

The full sky coverage and high-precision observations of *Gaia* offer the means of uncovering nearby UCDs through astrometric rather than purely photometric selection (Reylé 2018; Smart et al. 2019; Scholz 2020). *Gaia* provides a large homogeneous sample. The capability of *Gaia* to study the stellar to substellar boundary is illustrated in Sect. 5.2, where the luminosity function can be computed for the first time with one unique dataset throughout the main sequence down to the brown dwarf regime. It nicely shows a dip in the space density at spectral type L3, defining the locus of the stellar to substellar boundary.

5.4.1. New UCD candidates in *Gaia* EDR3

As mentioned in Sect. 4.2, GCNS contains thousands of faint stars (WDs and low-mass stars) that have no parallax from *Gaia* DR2. We investigate the potential new UCD candidates in GCNS in more detail. Following the selection procedure from Reylé (2018), we selected UCD candidates from the M_G versus $G - J$ diagram (Fig. 24, left panel). GCNS contains 2879 additional candidates compared to *Gaia* DR2, 1016 of which have a median distance inside 100 pc. This is a valuable contribution to complete the solar neighbourhood census in the region

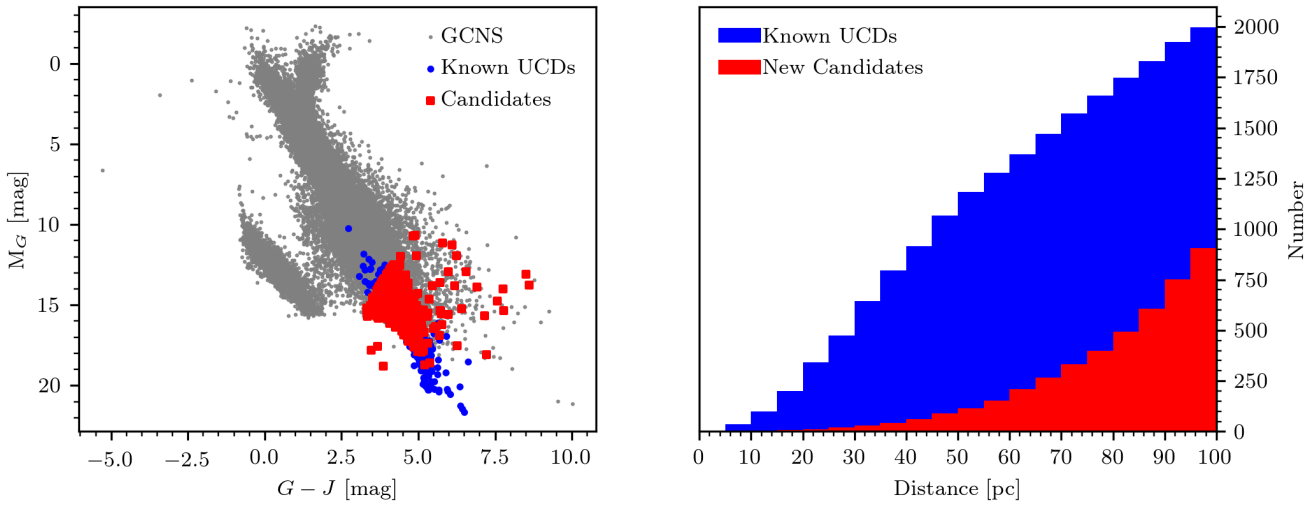

Fig. 24. *Left*: M_G vs. G–J diagram of stars in GCNS that are not found in *Gaia* DR2. The red dots are new UCD candidates, the blue points are known UCDs (spectral types between M7 and T8), and the grey points are the full GCNS sample. The new candidates are selected following the condition $M_G > -3 \times (G - J) + 25$, after removing stars whose probability of being a WD is higher than 20%. *Right*: distance distribution of the new candidates in the GCNS (red) and the known UCDs (blue).

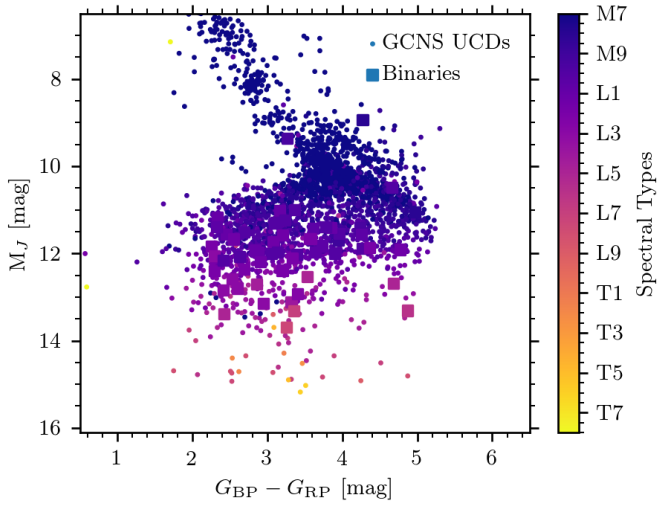

Fig. 25. CAMD of G_{BP}–G_{RP} [mag] against M_J [mag]. The full sample is from the GUCDS, and known binaries are over plotted as squares. Points are coloured by their published spectral types.

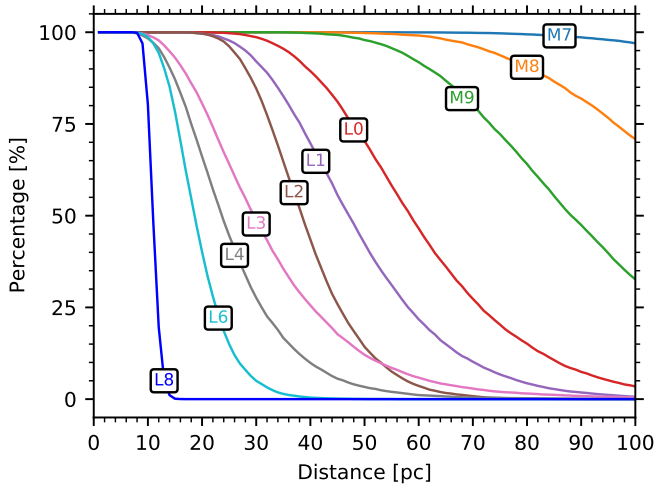

Fig. 26. Simulated completeness per parsec for each spectral type. Each spectral type from M7-L8 (right to left) is labelled next to its respective simulated completeness level. We skip L5 and L7 for better readability.

of the stellar to substellar boundary, as shown in the right panel of Fig. 24.

In Fig. 25 we examine G_{BP}–G_{RP} versus M_J for known UCDs taken from the *Gaia* Ultra-cool Dwarf Sample (Smart et al. 2017, 2019). The non-monotonic decrease of M_J with G_{BP}–G_{RP} indicates that G_{BP} is unreliable in the UCD regime, in agreement with the conclusions in Smart et al. (2019). For a full discussion and explanation of the limits on G_{BP}, see Riello et al. (2021).

5.4.2. GCNS completeness in the UCD regime

We show the simulated completeness for M7-L8 in Fig. 26. This was calculated using median absolute magnitudes M_G and standard deviations for each spectral type derived from the GCNS sample (in Sect. 4.2) and assuming a sky-isotropic G apparent magnitude limit of 20.4 mag with Monte Carlo sampling. Figure 26 indicates that an incompleteness begins at spectral type M7 and increases until L8, where the catalogue is only complete for the first 10 pc. The standard deviations of absolute magnitudes per spectral type bin are large (0.5–1 mag) and often have small sample sizes; therefore the noise in these simulations was quite large, which explains the crossing of the mean relation for some sequential spectral types.

5.4.3. UCD empirical completeness exceptions

We considered the simulated completeness from Fig. 26 with respect to a known sample, objects in the GUCDS identified in one of the *Gaia* releases, and spectral type from M7 to T6. This corresponds to 2925 sources. We find that 98 objects were not included in the GCNS that are in *Gaia* EDR3, but they either do not have parallaxes (34) or failed our probability selection (25), and 39 had parallaxes <8 mas. Of the 34 objects that did not have parallaxes, 21 did have parallaxes in *Gaia* DR2 but the five-parameter solutions in *Gaia* EDR3 were not published because their `astrometric_sigma5d_max` > 1.2 mas. This could be

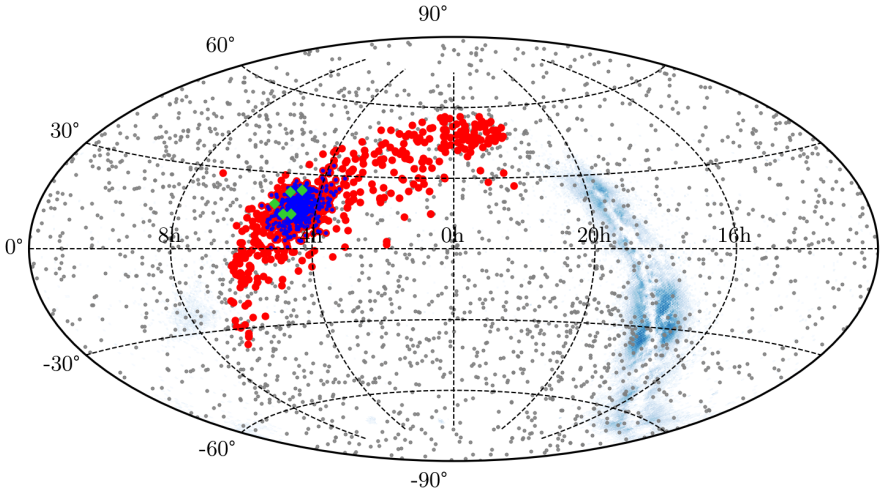

Fig. 27. Sky projection, in equatorial coordinates, of candidate Hyades members. The cloudy light blue structure in the background denotes the densest part of the Galactic plane in the direction of the Galactic centre. Grey dots denote all 3055 candidates, and filled red circles indicate the 920 sources that survived our ad hoc density filter aimed at suppressing contamination and bringing out the classical cluster and its tidal tails. Small blue dots denote 510 of the 515 Gaia Collaboration (2018a) members that are confirmed by *Gaia* EDR3 , and the green diamonds denote the five deprecated *Gaia* DR2 members.

because these objects are non-single or simply because they are very faint and at the limit of our precision.

An example of a system that we would expect to be in the GCNS is the nearby L/T binary Luhman 16 AB; *Gaia* DR2 5353626573555863424 and 5353626573562355584 for A and B, respectively, with $\pi = 496 \pm 37$ mas (Luhman 2013) and $G = 16.93$ & $G = 16.96$ mag. The primary is in *Gaia* DR2 and *Gaia* EDR3 (without complete astrometric solution in either), whilst the secondary is only in *Gaia* DR2 . This is a very close binary system with a short period, so that the use of a single-star astrometric solution may result in significant residuals that may have resulted in its exclusion in the current release.

5.5. Clusters within 100 pc

The 100 pc sample contains two well-known open clusters, the Hyades (Melotte 25, at ∼47 pc) and Coma Berenices (Melotte 111, at ∼86 pc). Both clusters stand out as density concentrations in 3D configuration as well as in 3D velocity space.

5.5.1. Membership

In order to identify candidate members, we largely followed the approach of Reino et al. (2018). Their method uses astrometry (positions, parallaxes, and proper motions), combined with radial velocity data when present, to compute 3D space motions and select stars as candidate members of each cluster. We slightly adopted the original approach and added an iterative loop in order to remove the dependence on the assumed initial conditions of the cluster. After convergence, the method attributes a membership probability to each star, expressing the statistical compatibility of the computed space motion of the star with the mean cluster motion, taking the full covariance matrix of the measurements as well as the cluster velocity dispersion into account (for details, see Reino et al. 2018). In contrast to Reino et al. (2018), who used the method on a limited-size field on the sky centred on the Hyades, we used the full all-sky GCNS catalogue. It is worth noting that the method only uses observables such as proper motions and radial velocities and does not depend on other parameters, in particular, on the GCNS probability (p) or the renormalised astrometric unit weight error (ruwe).

5.5.2. Hyades

For the Hyades, using the approach outlined above but limiting the radial velocities to those present in *Gaia* EDR3 (i.e. excluding ground-based values in GCNS), we identify 3055 candidate members. Their distribution on the sky (Fig. 27) shows three main features: (i) a dense concentration at the location of the cluster core, (ii) clear signs of two tidal tails, and (iii) a uniform spread of interlopers throughout the sky.

The tidal tails were discovered independently, based on *Gaia* DR2 data, by Meingast & Alves (2019) and Röser et al. (2019). These studies have noted the need to remove contamination, and both adopted a spatial density filter with subjective limits to highlight spatial over-densities: Meingast & Alves eliminated all sources with fewer than three neighbours within 20 pc, while Röser et al. first drew a sphere with a 10 pc radius around each star, then counted the number of stars that fell into this sphere, subsequently selected all spheres that were filled by six stars or more, and finally selected all stars that belonged to at least one of these spheres. In our sample, 920 candidate members remain after a density filter was adopted that somewhat arbitrarily accepted all stars with eight or more neighbours in a 10 pc sphere. By construction, the resulting set is strongly concentrated towards the classical cluster (630 stars are within two tidal radii, i.e. 20 pc of the centre) and the two tidal tails. The interpretation of the filtered sample clearly requires care, if only because edge effects are expected to be present close to the GCNS sample border at 100 pc.

The cluster was studied with *Gaia* DR2 data by Gaia Collaboration (2018a). We confirm 510 objects and refuse 5 of their 515 members. When the GCNS sample is extended to include the rejected stars with low probabilities ($p < 0.38$; Sect. 2.1) no members within 20 pc from the cluster centre are added. The closest "new" candidate member is found at 20.6 pc distance from the cluster centre and has $p = 0.01$. It also has an excessive ruwe = 2.73. This 19th magnitude object is most likely a partially resolved binary with suspect astrometry because a nearby polluting secondary star was detected but not accounted for in the *Gaia* DR3 data processing in about one-third of the transits of this object. We conclude that our membership, at least within two tidal radii, is not affected by the astrometric cleaning that underlies the GCNS sample definition.

Beyond the cluster core and surrounding corona, the candidate members show clear signs of tidal tails (Fig. 28). In the *Gaia* DR2 data, these tails were found to extend out to at least 170 pc from the cluster centre for the leading tail (which extends towards positive y into the northern hemisphere). Oh & Evans (2020) suggested that tail lengths of ∼400-800 pc can be expected. The two *Gaia* DR2 discovery studies used a sample

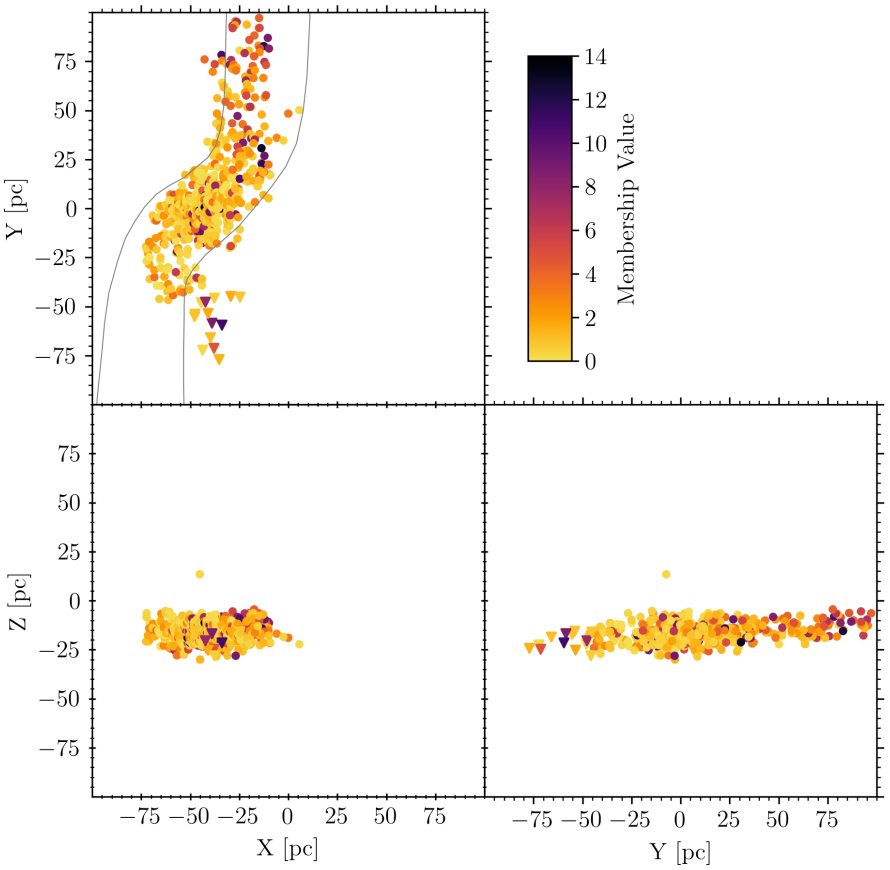

Fig. 28. Projections of the Hyades and its tidal tails in Galactic Cartesian coordinates (X, Y, Z) with the Sun at the origin. The grey lines (courtesy of Stefan Meingast) denote the approximate contours of the Hyades tidal tails as simulated by Chumak et al. (2005). The trailing tail shows a peculiar bend (triangles) where stars deviate from the simple N-body model prediction. These stars are well compatible with the cluster's space motion and agree well with the remaining cluster population based on their location in the CAMD (Fig. 29). We have no reason to assume that they do not belong to the cluster.

out to 200 pc from the Sun. Because the GCNS sample is by construction limited to a distance of 100 pc from the Sun, we cannot use GCNS to study the full extent of the tails. The GCNS does confirm, however, the *Gaia* DR2 -based observation that the trailing tail (at southern latitudes and negative y values) is less pronounced, deviates (triangles in Fig. 28) from the expected S-shape predicted by N-body simulations (e.g. Chumak et al. 2005; Kharchenko et al. 2009), and is currently dissolving (see also Oh & Evans 2020).

The classical G_{BP}–G_{RP} CAMD that is displayed in the left panel of Fig. 29 shows a narrow main sequence that extends the Gaia Collaboration (2018a) sequence based on *Gaia* DR2 by ~2 magnitudes towards fainter objects, a well-defined white dwarf sequence, and a clear sign of an equal-mass binary sequence that extends to the faintest objects ($M_G \sim 15$ mag). Further noticeable features in the CAMD include the broadening of the main sequence for M dwarfs, caused by radius inflation (e.g. Jaehnig et al. 2019), and a hook at the faint end comprising ~50 low-mass objects. The latter feature has been present as an artefact in *Gaia* DR2 (e.g. Lodieu et al. 2019) and is caused by spurious mean G_{BP} magnitudes exhibited by faint red targets for which negative G_{BP} transit fluxes that remain after background subtraction were not accounted for while forming the mean published G_{BP} magnitudes. This hook entirely consists of objects $G_{BP} < 20.3$, which would therefore be cut had we applied the photometric quality filter suggested in Riello et al. (2021).

As expected, the G-G_{RP} CAMD in the right panel of Fig. 29 shows a continuous, smooth main sequence all the way down to $M_G \sim 17$ mag. This CAMD shows another cloud of ~20 outliers to the right above the main sequence. These objects have problematic G_{BP} and G_{RP} magnitudes, as indicated by their non-nominal BP/RP flux excess values. These sources can be identified using the blended fraction β as described in Riello et al. (2021). Because G_{BP} and G_{RP} are biased in the same way for these sources (because more than one source lies within the BP/RP windows, which has not been accounted for in the *Gaia* EDR3 processing), the difference G_{BP}–G_{RP} is fairly accurate. The CAMD outliers in both the G_{BP}–G_{RP} version (hook) and the G–G_{RP} version are fully explained by known features of the *Gaia* EDR3 photometry and are not correlated to the position of the stars in the cluster or tidal tails. All in all, both CAMDs demonstrate the overall exquisite (and improved) quality of the *Gaia* EDR3 astrometry and photometry.

5.5.3. Coma Berenices

The 100 pc sample contains a second open cluster, Coma Berenices. It has similar age (~800 Myr) and tidal radius (~7 pc) as the Hyades, but is twice as distant, close to the GCNS sample limit. The cluster has been studied with *Gaia* DR2 data by Gaia Collaboration (2018a), who found 153 members in a limited-size field of view. We used the *Gaia* EDR3 astrometry and repeated the same procedures as outlined above. Within the central 14 pc we confirm 146 of the *Gaia* DR2 members and add 15 new candidate members. Tang et al. (2018) noted that the cluster is elongated along its orbit towards the Galactic plane, and subsequently reported tidal tails (Tang et al. 2019). Our *Gaia* EDR3 candidate members show very clear signs of tidal tails beyond two tidal radii from the cluster centre, but their precise shape and membership depends sensitively on the spatial density filter that is needed to remove contamination from the all-sky GCNS sample. Moreover, a study of the cluster and its tidal tails based on the GCNS sample is complicated because it lies close to the sample border.

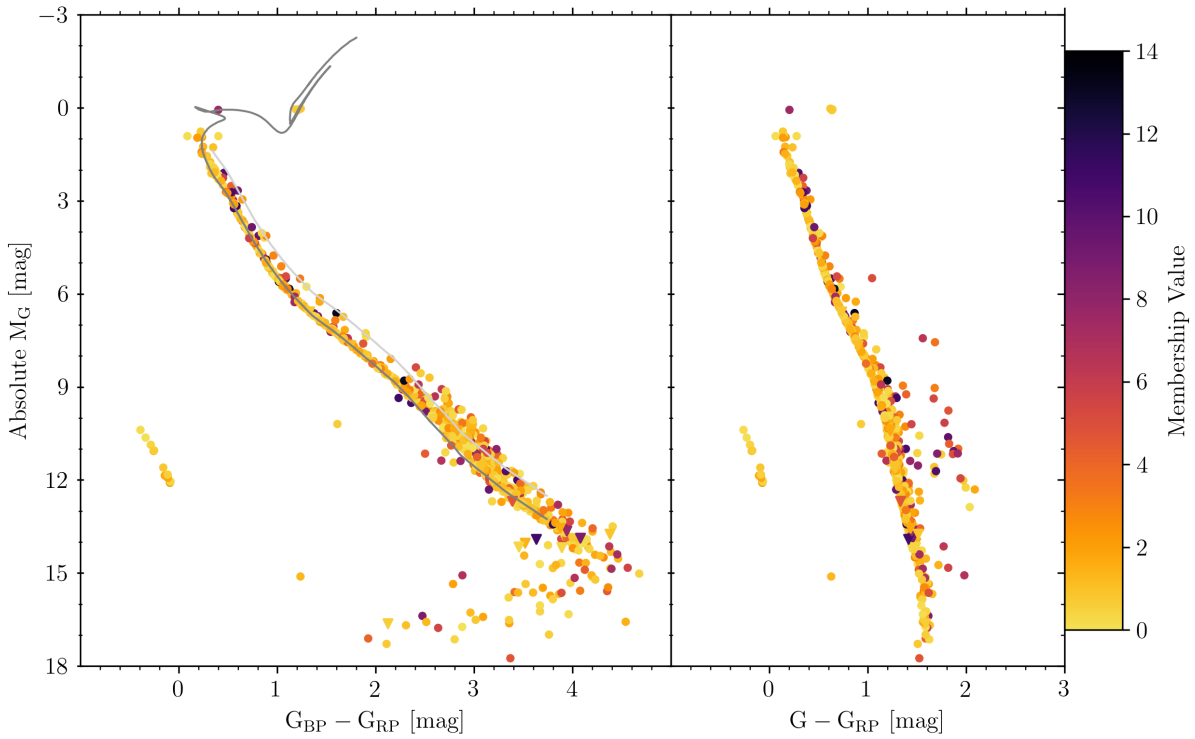

Fig. 29. CAMDs for the 920 Hyades candidate members using G_{BP}–G_{RP} (*left*) and G–G_{RP} (*right*). Extinction and reddening are not included but are generally negligible for the Hyades. Absolute magnitudes have been computed using `dist_50` as distance estimate. The hook at the faint end of the G_{BP} -G_{RP} main sequence (*left panel*) is a known artificial feature of *Gaia* EDR3 caused by spurious G_{BP} magnitude estimates for very faint intrinsically red sources. The outliers to the right above the G-G_{RP} main sequence (*right panel*) have biased G_{RP} and G_{BP} magnitudes, as indicated by the high BP/RP flux excess values. Because their G_{BP} and G_{RP} magnitudes are biased by the same amount, the G_{BP}–G_{RP} value (*left panel*) is fairly accurate. The colour of the symbols encodes our membership probability, with low c values (yellow) indicating highly probable members. The grey curves in the *left panel* denote a 800-Myr PARSEC isochrone and its associated equal-mass binary sequence (both based on *Gaia* DR2 passbands); they are not a best fit, but are only meant to guide the eye. Fourteen stars are marked with triangles; they are nearly indistinguishable from the remaining stars in the two CAMDs. These correspond to a group of stars that deviates from a simple N-body prediction for the development of the tidal tails, see the (x, y)-diagram in Fig. 28.

5.6. Stellar multiplicity: resolved systems

Statistical studies of stellar multiplicity are key to a proper understanding of many topics in modern-day astrophysics, including star formation processes, the dynamics of dense stellar environments, the various stages of stellar evolution, the formation and evolution of planetary systems, the genesis of extreme high-energy phenomena (supernovae, gamma ray bursts, and gravitational waves), and the formation of the large-scale structure of the Universe (e.g. Belokurov et al. 2020, and references therein). Early investigations of the statistical properties of stellar systems in the solar neighbourhood based on small sample sizes (hundreds of stars, e.g. see Duquennoy & Mayor 1991; Raghavan et al. 2010, and reference therein) have revealed not only a high binary (and higher multiplicity) fraction, but also trends in stellar multiplicity with spectral type, age, and metallicity. With the improvements in the host of techniques used to search for stellar binaries, which include photometry, spectroscopy, astrometry, high-contrast imaging, and interferometry (e.g. Moe & Di Stefano 2017), such trends are now being placed on solid statistical grounds based on typical sample sizes of tens of thousands of systems (e.g. El-Badry & Rix 2018; Tokovinin 2018; Moe et al. 2019; Merle et al. 2020; Price-Whelan et al. 2020). A new revolution is in the making, however, with the *Gaia* mission bound to provide a further order-of-magnitude increase in known stellar systems across all mass ratios and orbital separations (Söderhjelm 2004, 2005), detected based on the astrometric, spectroscopic, photometric, and spatial resolution information of *Gaia*.

The first two *Gaia* data releases (DR1 and DR2) have permitted detailed investigations with unprecedented precision of the regime of spatially resolved intermediate- to wide-separation binaries (e.g. Andrews et al. 2017; Oh et al. 2017; Oelkers et al. 2017; El-Badry & Rix 2018; Moe et al. 2019; Jiménez-Esteban et al. 2019; Hartman & Lépine 2020, and references therein). Such systems are of particular interest because of their low binding energies, they can be used as probes of the dynamical evolution history of the Galaxy and of the mass distribution and number density of dark objects in the Milky Way (e.g. El-Badry & Rix 2018, and references therein; Hartman & Lépine 2020, and references therein). They were born at the same time and in the same environment but evolved in an entirely independent way, therefore they are very useful tools for testing stellar evolutionary models, they can be used as calibrators for age and metallicity relations (e.g., Jiménez-Esteban et al. 2019, and references therein). Because they are common in the field (Raghavan et al. 2010), they are natural laboratories in which to study the effect of stellar companions on the formation, architecture, and evolution of planetary systems (e.g. Desidera & Barbieri 2007; Deacon et al. 2016; Kraus et al. 2016; Kane et al. 2019).

Using the updated astrometric information in the GCNS catalogue, we performed a new search for wide binaries within 100 pc of the Sun. We first identified neighbouring objects with

Table 3. Summary data on binary pairs in the GCNS catalogue of wide binaries.

SourceId 1 (Primary)	SourceId 2 (Secondary)	Separation (arcsec)	ΔG	Proj. Sep. (au)	Bound	Hyades	Coma	Binary
83154862613888	83154861954304	3.8353	3.2631	244.7406	true	false	false	true
554329954689280	554329954689152	3.7164	0.4823	358.7470	true	false	false	true
1611029348657664	1611029348487680	6.1252	0.8744	513.6358	true	false	false	true
1950331764866304	1962117155125760	9.3117	6.8372	810.1125	false	false	false	true
...	...	...	...	...	...	...	...	...

Notes. The full table is available at the CDS.

an angular separation in the sky <1 deg (which implies a non-constant projected separation in au), similarly to Hartman & Lépine (2020). We did not impose a lower limit on the projected separation in order to characterise the loss in efficiency in detecting pairs when the resolution limit of *Gaia* EDR3 is approached. We then followed Smart et al. (2019) and adopted standard criteria to select a sample of likely bound stellar systems: (1) scalar proper motion difference within 10% of the total proper motion ($\Delta\mu < 0.1\mu$), and (2) parallax difference within either 3σ or 1 mas, whichever is greater ($\Delta\hat{\varpi} < \max[1.0, 3\sigma]$). We further refined the selection with a second pass based on the requirement of boundedness of the orbits, following El-Badry & Rix (2018), but placing the more stringent constraint $\Delta\mu < \Delta\mu_{\rm orbit}$, with $\Delta\mu_{\rm orbit}$ defined as in Eq. (4) of El-Badry & Rix (2018).

The application of our selection criteria to the GCNS catalogue allowed us to identify a total of 16 556 resolved binary candidates (this number increases to 19 176 when we do not impose the bound orbit criterion). The relevant information is reported in Table 3. The selection by construction contains objects that are co-moving because they are members of rich open clusters (Hyades and Coma Berenices), more sparsely populated young moving groups (e.g. Faherty et al. 2018), as well as higher-order resolved multiples in which more than one companion is identified to either member of a pair. In Table 3 we flag both higher-order multiples and cluster members (1758 and 286, respectively) based on the updated cluster membership list in Sect. 5.5.

The upper left panel of Fig. 30 shows the colour-magnitude diagram for the primaries in the 100 pc wide-binary candidate sample. The plot is colour-coded by magnitude difference with the secondary. A small number of objects are removed as they do not have full colour information in *Gaia* EDR3. The diagram is almost free of spurious objects located in between the main sequence and the WD cooling sequence, which amount to no more than 0.2% of the sample. These objects are likely misclassified due a variety of reasons that are summarised in Hartman & Lépine (2020). Similarly to El-Badry & Rix (2018) and Hartman & Lépine (2020), for instance, the diagram also displays an indication of a secondary main sequence, offset upward by ~0.5 mag particularly in the $1.0 \lesssim (G_{\rm BP}-G_{\rm RP}) \lesssim 2.0$ mag range. This is the unresolved binary sequence composed of hierarchical systems in which one or both of the resolved components is itself a spatially unresolved binary with a typical mass ratio $q \gtrsim 0.5$ (e.g. Widmark et al. 2018, and references therein).

Further evidence of the presence of the photometric binary main-sequence is found by colour-coding the plot in the upper left panel of Fig. 30 using the value of the ruwe, which exhibits a notable excess in this region (plot not shown, but see e.g. Belokurov et al. 2020). Overall, ~24% of the objects in our catalogue have ruwe $\gtrsim$1.4 (indicative of an ill-behaved astrometric solution), and in ~2% of the cases, both components of a binary have a high ruwe value. These numbers might be explained based on the combined effects from higher-order multiples with short-period components (this number is difficult to derive as it entails understanding the selection function of short-period binaries with wide-separation stellar companions) and larger samples of intermediate-separation binaries that become unresolved or partially resolved as a function of increasing distance (preliminary estimates indicate that this percentage is about 15–20%).

The upper right panel of Fig. 30 shows the G mag difference ΔG of our wide-binary candidates as a function of angular separation. The sample of objects flagged as Hyades and Coma Ber cluster members, as determined in Sect. 5.5, is also reported. The slope of increasingly lower ΔG at separations <10″ is the footprint of the *Gaia* sensitivity loss, which nicely follows the behaviour in contrast sensitivity shown in Fig. 9 At separations $\gtrsim$10″, the interval of ΔG is essentially independent of separation. Interestingly, all Coma Ber bona fide cluster members flagged as candidate binaries reside at very wide separations (at the distance of Coma Ber, the typical projected separation $\gtrsim 4 \times 10^4$ au). Even when the requirement of formally bound orbits is enforced, a significant fraction of the very wide binaries could still be a result of chance alignment. We estimated the contamination rate of our sample of wide binaries using the GeDR3 mock catalogue (Rybizki et al. 2020). The catalogue does not contain any true binaries: an adoption of our selection criteria for pair identification in the mock catalogue provides a direct measurement of the number of false positives (pairs due to chance alignment) in our sample, particularly in the regime of very wide separations. When we applied our two-pass search criteria, the mock catalogue returned five pairs, all at a separation $\gtrsim$1000 au. This means that the contamination level in our sample probably is 0.05–0.1%.

A comparison with the recent DR2-based catalogue of wide binaries produced by El-Badry & Rix (2018) shows good agreement with our selection in the overlapping regime of separations when cluster members and higher-order multiples are excluded (see Fig. 30, bottom left panel). When we restrict ourselves to the regime of projected physical separations defined by El-Badry & Rix (2018), our candidates match those found by El-Badry & Rix (2018) within 100 pc to 78.1%. The discrepancy is likely due to significantly revised values of parallax and proper motion in *Gaia* EDR3 with respect to the DR2 values. We also note an overall increase of ~20% in detected pairs with respect to the El-Badry & Rix (2018) 100 pc sample. The larger number of close pairs identified in the GCNS samples of candidates is a possible indication of a moderate improvement in sensitivity in the ≈1″–3″ regime with respect to DR2-based estimates (e.g. Brandeker & Cataldi 2019).

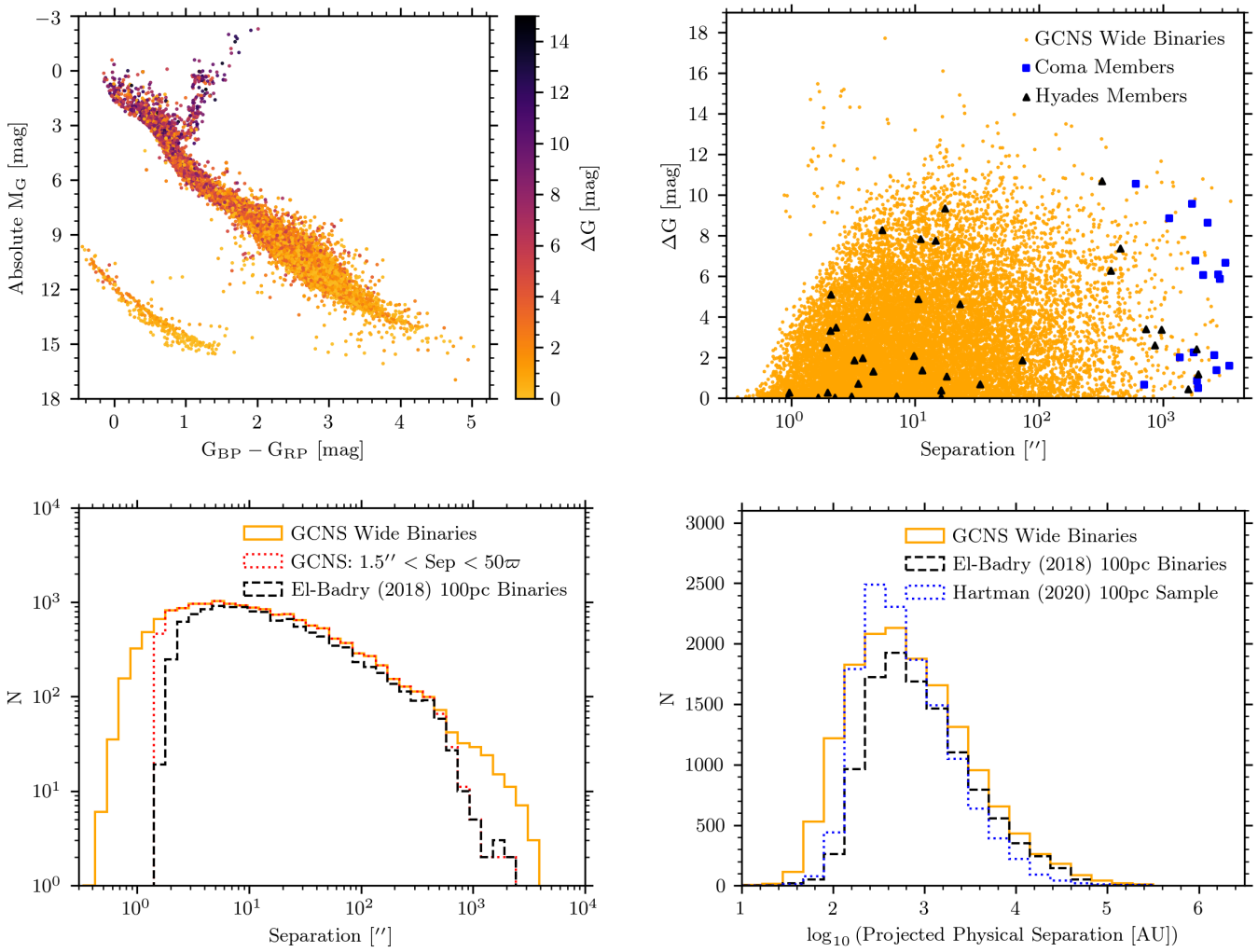

Fig. 30. *Top left*: CAMD for the GCNS systems colour-coded by the magnitude differences of the binary components. *Top right*: separation vs. *G* mag difference for the resolved stellar systems in GCNS (orange points). Known members of the Hyades (black triangles) and Coma Ber (blue squares) clusters are highlighted. *Bottom left*: histogram (solid orange) of separations for wide binaries in the GCNS sample compared to the DR2-based catalogue (dashed black) from El-Badry & Rix (2018). The dotted red histogram corresponds to the separation distribution of GCNS wide-binary candidates adopting the exact boundaries in El-Badry & Rix (2018). *Bottom right*: physical projected separation distribution for the wide-binary candidates identified in this work (solid orange) compared to those from El-Badry & Rix (2018) (dashed black) and Hartman & Lépine (2020) (dotted blue), restricted to systems within 100 pc.

The bottom right panel of Fig. 30 shows the projected physical separation of our wide-binary candidates, compared to the distributions of the same quantity in the El-Badry & Rix (2018) and Hartman & Lépine (2020) catalogues, both restricted to $d < 100$ pc (and the latter with Bayesian binary probability >99%). All distributions peak around $10^{2.5}$ au, and they all exhibit the same exponential decay at wider separations. Finally, we retrieve 25 of the 63 very wide binaries within 100 pc in the Jiménez-Esteban et al. (2019) catalogue (this number increases to 41 if we lift the requirement on formally bound orbits). Similarly to the searches performed by El-Badry & Rix (2018) and Hartman & Lépine (2020), we find no evidence of bi-modality in the distribution of projected physical separations due to a second population of binaries with companions at >100 000 au, as had been previously suggested. As a matter of fact, such a feature is instead clearly seen (plot not shown) in our first-pass sample selected without imposing that the orbits be physically bound.

Using the spectral type and median M_G calibration found in Sect. 4.2 and the list of binary candidates cleaned for cluster members and higher-order multiples, we briefly comment on the wide-binary fraction f_{WB} in the 100 pc sample. For instance, we obtain $f_{WB} = 4.8^{+0.4}_{-0.3}\%$ (with $1-\sigma$ errors derived using the binomial distribution, e.g. see Burgasser et al. 2003; Sozzetti et al. 2009) for M dwarfs within 25 pc in the regime of separations >2″ (171 wide systems in a sample of 3555 M-type stars). This differs at the $\sim 3.5\sigma$ level from the 7.9 ± 0.8 multiplicity rate reported by Winters et al. (2019), although subtracting the approximately 25% of higher-order multiples from their sample the results become compatible within 1.6σ. As highlighted by the histogram in Fig. 31, in the volume-limited 100 pc sample the wide-binary fraction appears constant for F- and G-type dwarfs, with a measured rate entirely in line with previous estimates ($f_{WB} \simeq 10-15\%$) in the literature (e.g., Moe et al. 2019, and references therein). The hint of a decline in wide-binary fraction for K-dwarfs is likely real, as based on the spectral type versus M_G relation provided in Sect. 4.2 we are complete for all K types. The clear decline in f_{WB} for the M dwarf sample is real, and only mildly affected by incompleteness at the latest sub-spectral types (>M7).

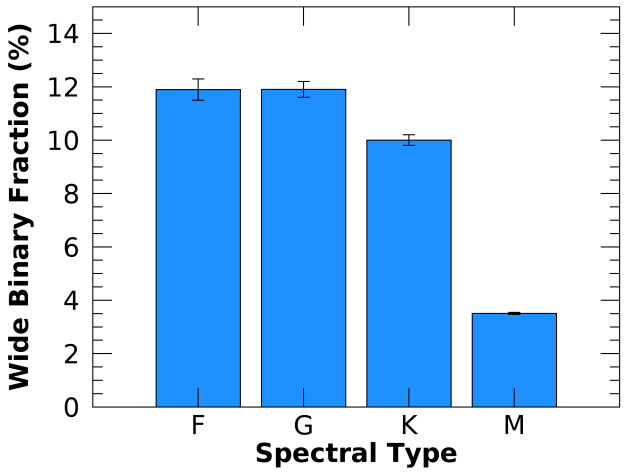

Fig. 31. Histogram of the wide-binary fraction within 100 pc as recovered from the GCNS catalogue. Error bars, representing 1σ confidence intervals, are derived from a binomial distribution.

5.7. Stellar multiplicity: unresolved systems

The advent of precise photoelectric photometry in the latter half of the last century contributed to the discovery of a significant level of close binarity amongst the stellar populations. Photometric observations in coeval populations with very low dispersion in chemical composition (e.g. rich clusters) reveal a faint sequence parallel to the locus of the main sequence in the CAMD – e.g. Stauffer (1984) and references therein. Given sufficient precision in the photometric measurements a single star sequence is often accompanied by a second sequence at brighter magnitudes and redder colours. The reason is that in a significant fraction of spatially unresolved binaries, twins (i.e. equal-mass binaries) show a vertical elevation of 0.75 mag in the CAMD, while extreme mass ratio binaries exhibit a significantly redder colour close to the same brightness as a single star (see previously in Sect. 5.6 and Fig. 30), and an elevated locus of unresolved binaries is populated by all mass ratios between these two extremes. The high-precision photometry of *Gaia* leads to particularly fine examples of this phenomenon in clean astrometric samples of cluster stars (e.g. Gaia Collaboration 2018a).

Two large coeval populations of stars overlap in the GCNS sample: the Hyades and Coma Berenices clusters (Sect. 5.5). The Hyades in particular present a rich sequence of photometrically unresolved binaries that is evident in Fig. 29. Using the cluster members derived in Sect. 5.5, and limiting our selection to within two tidal radii of the respective cluster centres, we made subsamples of the GCNS catalogue for the Hyades and Coma Ber. The only additional filtering on photometric quality applied in this case was as defined in Gaia Collaboration (2018a), namely $\sigma_G < 0.022$ and $\sigma_{BP,RP} < 0.054$, along with their photometric quality cut (via `phot_bp_rp_excess_factor`; see Appendix B in Gaia Collaboration 2018a). We traced the locus of the single-star sequence using a low-order polynomial fit that allowed us to subtract the slope in the CAMD. This yielded a set of ΔG versus colour. Marginalising over the whole range in colour and employing a two-component Gaussian mixture resulted in the models for the star counts versus ΔG shown in Fig. 32.

The difference in ΔG between these two components is ~0.7 mag, as expected for a dominantly single-star population along with a subordinate population of near equal–mass but unresolved binaries. According to this simple model the binary

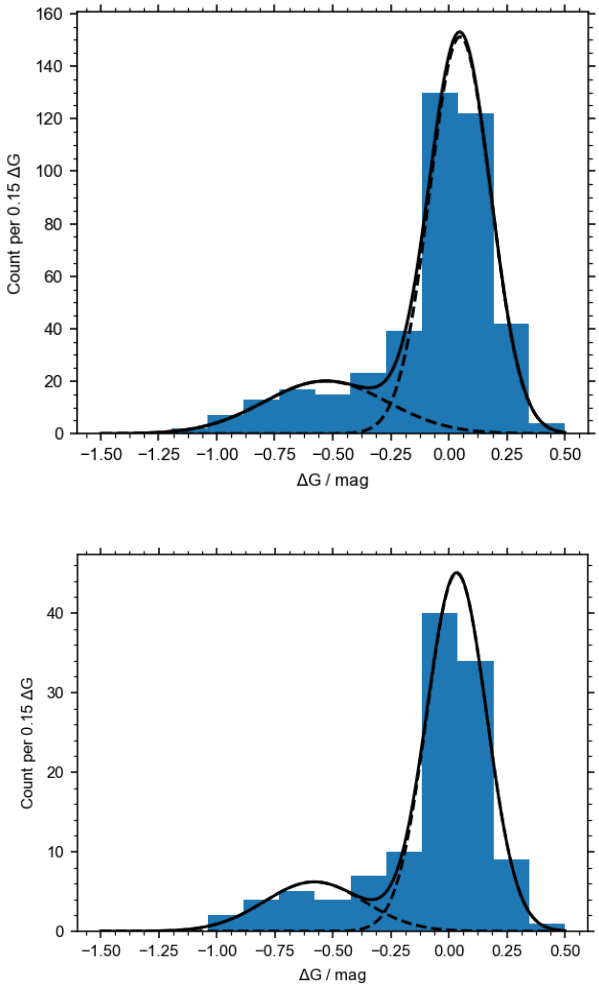

Fig. 32. Two-component Gaussian mixture model of the distributions of star counts per 0.15 mag ΔG bin for the Hyades cluster in the *upper panel* and Coma Ber in the *lower panel*, as described in the text.

fraction, measured as the ratio of weights of the subordinate to dominant component and counting one star in the latter and two stars in the former, is 34% for the Hyades and 31% in Coma Ber. This is for the range $0.5 < G_{BP} - G < 2.5$, which corresponds roughly to main-sequence masses in the range $1.4\,M_\odot$ down to $0.2\,M_\odot$ (according to simulations – see below).

The general field population sampled by the GCNS is neither coeval nor chemically homogeneous. However, an analogous procedure can be applied in its CAMD, noting that the ghostly signature of unresolved binarity is easily visible at intermediate colours (e.g. the top right panel in Fig. 13). Furthermore, it is instructive to apply the same procedure to *Gaia* CAMD simulations generated without binaries, a fiducial level of binarity, and a few mid-fractions.

In GOG, in particular, those provided as part of *Gaia* EDR3 (Gaia Collaboration 2021), binary stars are generated but the unresolved binaries do not have the fluxes of the combined components. They have the flux of the primary. Therefore the sequence of twin binaries is not present in GOG, and we used another set of simulations to analyse unresolved binaries. For this we used the last version of the Besançon Galaxy model, where the initial mass function and star formation history were fitted to *Gaia* DR2 data (see Mor et al. 2018, 2019, and references therein) where the star formation history of the thin disc is assumed to decrease exponentially. The stellar evolutionary tracks we used

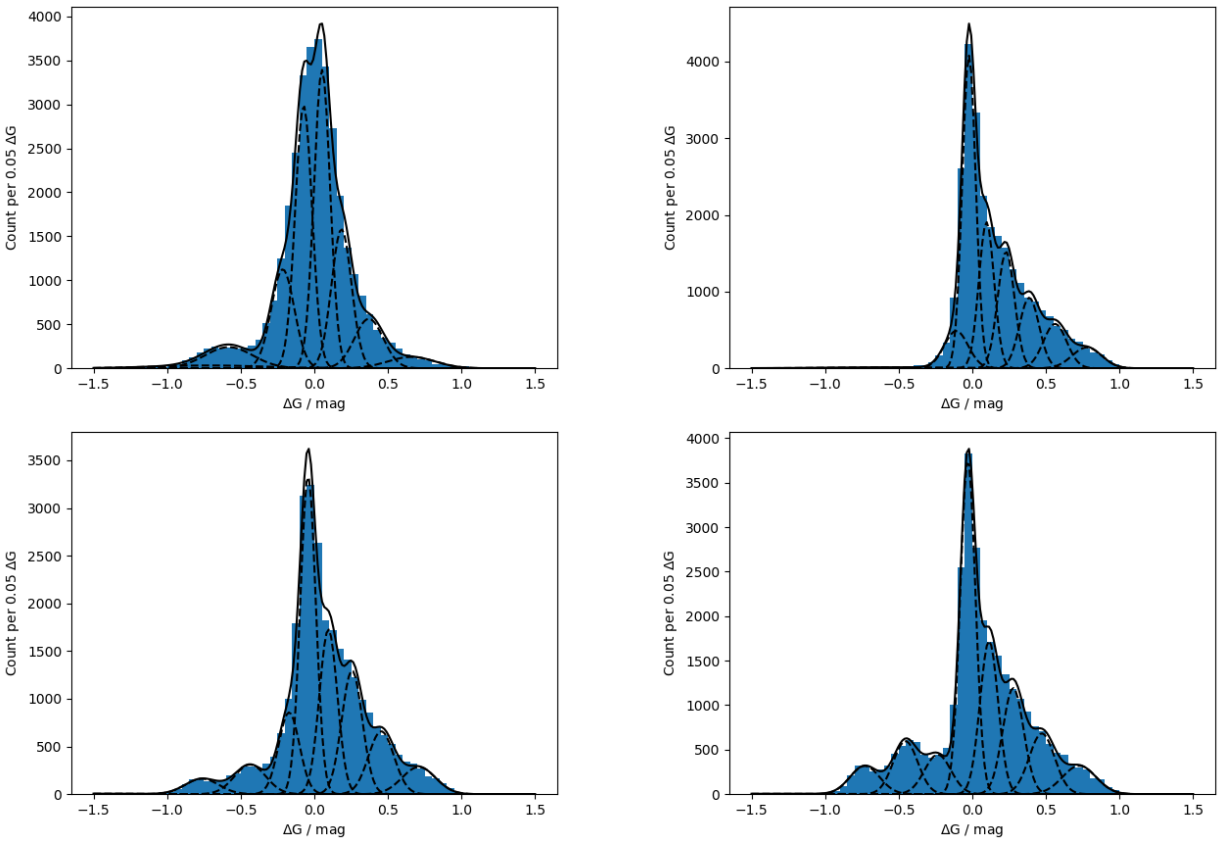

Fig. 33. Gaussian mixture models of the distribution of star counts per 0.1 mag ΔG bin in the range $0.5 < G_{\rm BP} - G < 1.0$ after the slope of the main sequence is subtracted in the CAMD. *From top to bottom, first panel*: GCNS, *next three panels*: simulations employing no binarity, half–fiducial, and fiducial binarity according to the prescription of Arenou (2011).

Table 4. Component weights contributing to the Gaussian mixture models for ΔG indicating photometrically unresolved binarity.

	Observations	Arenou (2011)	×0.8	×0.5	×0.0
Component weights	−0.716 : 0.0180	−1.440 : 0.0044	−1.331 : 0.0054	−2.157 : 0.0008	−2.000 : 0.0019
above main sequence	−0.580 : 0.0579	−0.730 : 0.0527	−0.713 : 0.0476	−1.393 : 0.0028	−0.557 : 0.0105
$0.5 < G_{\rm BP} - G < 1.0$	−0.214 : 0.1301	−0.454 : 0.0823	−0.424 : 0.0871	−0.763 : 0.0322	
(ΔG : weight)		−0.235 : 0.0660		−0.436 : 0.0573	
TOTAL WEIGHT:	0.2060	0.2054	0.1401	0.0931	0.0124

Notes. Columns 3 to 6 are from simulations using the prescription of Arenou (2011) in full and at fractional reductions of 0.8, 0.5, and 0.0.

are the new set from the STAREVOL library (see Lagarde et al. 2017, and references therein). The complete scheme of the model is described in Robin et al. (2003), while the binarity treatment is explained in Czekaj et al. (2014). The generation of binaries, probability, separation, and mass ratio is the same as in the GOG simulations. However, in contrast to the GOG simulations, unresolved binaries are treated such that the magnitude and colours reflect the total flux and energy distribution of the combined components.

This allows a comparison with the cluster results above and also provides a confirmation of the level of realism in the simulations. The binary angular separation s in arcseconds at which simulated pairs become unresolved was assumed to follow Eq. (2) in Sect. 4.1.3. The fiducial level of binarity follows Arenou (2011) (see also Arenou 2010), which is near 60% for solar mass stars and decreases to about 10% for stars of mass 0.1 $M_\odot$. Figure 33 shows Gaussian mixture models for histograms of ΔG for the stellar main sequence in the colour range $0.5 < G_{\rm BP} - G < 1.0$ after subtracting the sloping locus in the CAMD, this time tracing the latter using the dominant component in a Gaussian mixture model in colour bins of 0.05 mag; a final mixture model was again employed for the resulting marginal distributions of star counts versus ΔG. Table 4 quantifies the measured level of photometrically unresolved binarity by summing the weights of all components that significantly contribute to the counts for the range in ΔG that is affected by unresolved binarity according to the simulations. The observations appear to match the fiducial binarity level simulation by this metric very well.

Given the approximations and assumptions made in this simple analysis, the agreement between simulations and observations is gratifying. A significant source of uncertainty on the observational side is the assumed angular resolution. In reality, the extant processing pipeline (at Data Reduction Cycle 3

corresponding to *Gaia* EDR3) does not deblend close pairs observed in single transit windows (Gaia Collaboration 2021). The effects on photometry and astrometry depend on the scan angle with respect to the position angle of the binary in a given transit, as well as on the angular separation. Another limitation on the simulation side is the treatment of the distribution in metallicity, as shown by the sharper features in the histogram counts in Fig. 33. Because of these complications, we draw no more quantitative conclusions as to the true level of binarity in the GCNS sample. Further and more detailed studies of the effects of binarity (e.g. Belokurov et al. 2020; Laithwaite & Warren 2020) are clearly warranted but are beyond the scope of this demonstration work.

5.8. White dwarfs

5.8.1. White dwarf selection

To recover the WD population in our catalogue, we started analysing all 1 040 614 sources with $\hat{\varpi} > 8$ mas for which the three *Gaia* photometric passbands are available. We used the 29 341 sources from this larger sample that are in common with three different catalogues of known WDs (Gentile Fusillo et al. 2019, Torres et al. 2019, and Jiménez-Esteban et al. 2018) to build training and test datasets for our WD random forest classification algorithm. We selected 20 000 of them to constitute the WD sample in the training dataset, and the other 9341 sources became the test dataset.

After these known WDs were excluded from the whole set of 1 040 614 sources with $\hat{\varpi} > 8$ mas and *Gaia* photometry, we randomly selected 40 000 and 37 364 sources to constitute the training and test dataset of non-WD sources, respectively. We chose these particular numbers of sources in order to use a training dataset with twice the number of non-WDs with respect to the number of WDs, but four times for the test dataset. This is useful to detect whether the ratio of WDs in the sample analysed affects the classification. This random selection of non-WDs was made with the aim of maintaining the colour distribution of the whole sample and better populating our sample with non-WDs with blue colours that might be confused with WDs. Thus, we selected 9.6% of the sources having $G - G_{\rm RP} < 0.75$ mag and the rest with larger $G - G_{\rm RP}$ values. This selection resulted in a set of 60 000 training and 46 705 test sources. Their distribution in the CAMD for the training dataset is shown in Fig. 34 (the CAMD diagram for the test dataset looks very similar). These figures show the concentration of WDs in $G - G_{\rm RP} < 1$ mag, increasing the normalised colour distribution in this range. As explained, this was the main reason to better populate blue colours in the non-WD dataset to avoid confusion with the WD sample.

Using these datasets, we trained the random forest algorithm with the purpose of classifying WDs. We used the Python random forest classifier, performing a cross-validation to obtain the most appropriate hyperparameters[12]. As information for the classification of WDs we considered the three photometric *Gaia* magnitudes ($G, G_{\rm BP}$, and $G_{\rm RP}$), proper motions (μ_α and μ_δ), and parallaxes ($\hat{\varpi}$), and also their uncertainties.

The most important features that help in the WD classification are the red magnitude (24.7% of the total weight) and the parallax uncertainty (14.5%). The parallax uncertainty is more important than the parallax itself, which is, in fact, one of the least important parameters because WDs are well separated from non-WDs in a CAMD inside 125 pc.

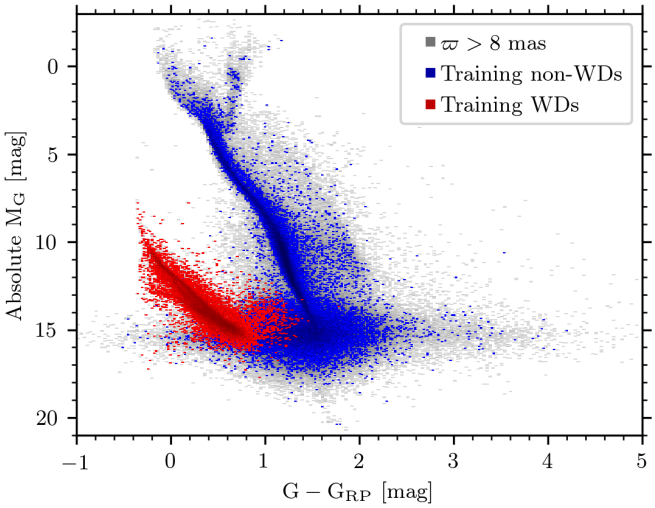

Fig. 34. CAMD for the training dataset based on which we classified the WDs. Red data points represent the WD population. Grey points are all sources in the whole $\hat{\varpi} > 8$ mas sample from which these samples were extracted. The appearance of the test dataset is quite similar to the one plotted here.

After the algorithm was trained, we evaluated how well the test dataset was classified. It correctly classified 9160 WDs (98.1% of the total) and 37 214 non-WDs (99.6% of the total) in the test dataset. The resulting list of WD candidates is contaminated by 147 non-WDs (representing 1.6% of the list of WD candidates derived from the test dataset). We then applied the classification algorithm to the whole $\hat{\varpi} > 8$ mas dataset with three passband *Gaia* photometry. The random forest algorithm outputs a value representing the probability of each source of being a WD.

From the total of 1 040 614 sources we found 32 948 sources with a probability of being a WD ($P_{\rm WD}$) higher than 0.5[13]. After the selection in Sect. 2, 21 848 of these sources were included in GCNS dataset[14]. The CAMD of these WD candidates in GCNS is shown in Fig. 35. We verified that the distribution in the sky of the WD candidates is homogeneous, as it is expected to be in the 100 pc bubble. Of the 11 106 sources with $P_{\rm WD} > 0.5$ having $\hat{\varpi} > 8$ mas that are not included in GCNS catalogue, only 815 with dist_1<0.1 fail the GCNS p criteria. They are at the red side of the WD locus, where more contamination is expected, and it is therefore possible that they are not real WDs.

When we compared our list of WD candidates with the 29 341 WDs used for training and testing the algorithm (extracted from Gentile Fusillo et al. 2019, Torres et al. 2019, and Jiménez-Esteban et al. 2018), we detected 2553 new WD candidates that were not included in the referenced bibliography (see Fig. 36 to see their position in the CAMD and their probability distribution of being a WD). These new candidates are mostly located in the red region of the WD locus, where the contamination is expected to be higher. In Fig. 37 we plot our WD candidates, $P_{\rm WD}$, and those of Gentile Fusillo et al. (2019), $P_{\rm GF}$.

There are 250 sources assigned $P_{\rm WD} > 0.5$ in this work, but which have a low probability in Gentile Fusillo et al. (2019).

[12] This cross-validation process returned the following best values for the hyperparameters: bootstrap = True, max_depth = 20, max_features = 'sqrt', min_samples_leaf = 4, min_samples_split = 10, n_estimators = 30.

[13] Full table available at the CDS.

[14] There are 7005 sources in GCNS that do not have all three *Gaia* photometric passbands. These sources were not assigned any value for the probability of being a WD.

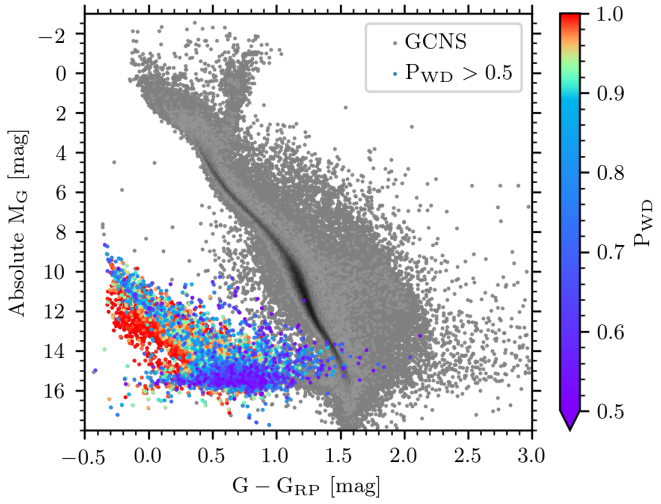

Fig. 35. CAMD of sources included in the GCNS (grey) and the WD candidates obtained with the random forest classification algorithm having $P_{WD} > 0.5$ (with the value of P_{WD} shown with the colour index).

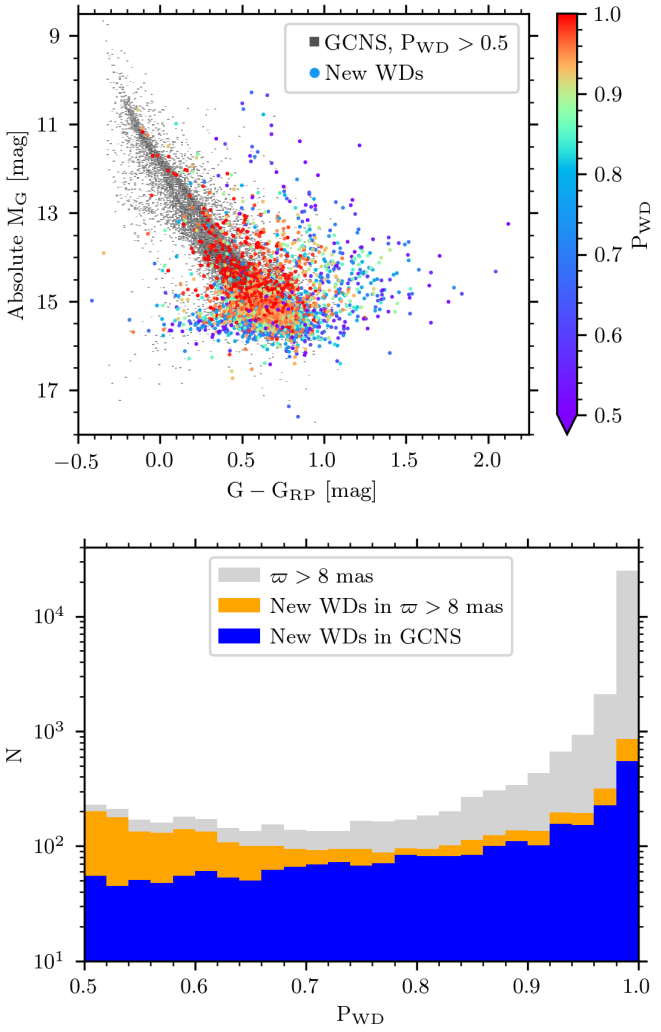

Fig. 36. *Top*: CAMD with the new WDs found here that were not included in Gentile Fusillo et al. (2019), Torres et al. (2019), or Jiménez-Esteban et al. (2018). *Bottom*: probability distribution of the new WD candidates.

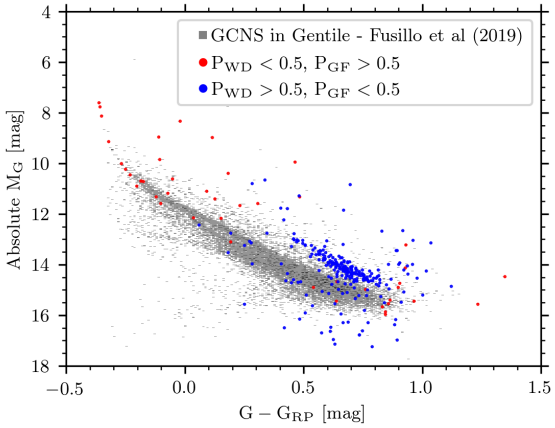

Fig. 37. Position in the CAMD of the WD candidates in Gentile Fusillo et al. (2019), P_{GF}, and our candidate P_{WD}. Coloured points indicate contradictory conclusions between the two studies.

Based on their position in the CAMD (blue points in Fig. 37), we suggest that this is probably due to very restrictive filtering in Gentile Fusillo et al. (2019) because these sources are mostly concentrated in the upper red part of the WD locus. On the other hand, 45 sources with $P_{WD} < 0.5$ are in our work, but are present in Gentile Fusillo et al. (2019) with $P_{GF} > 0.5$ (red points in Fig. 37). These red points include some sources that are located in the upper blue region of the WD locus, where our algorithm appears to fail to recognise these extreme sources as WDs. Some of these sources that are not recognised as WDs have very bright magnitudes compared with the training dataset we used (Sirius B and LAWD 37 are two examples of this). Because they are very few and are already contained in previous catalogues, they are easily recognisable. For completeness, we decided to include these 45 sources in a table available at the CDS including P_{GF} values.

5.8.2. White dwarf luminosity function

The white dwarf luminosity function (WDLF) tracks the collective evolution of all WDs since they were formed. Stars with masses up to $\sim 8\,M_\odot$ will become WDs, and their individual luminosity is determined by a relatively simple cooling law because all energy-generation processes have ceased, unless they are part of a binary system where later mass transfer can heat the envelope. In simple terms, the stored energy in the isothermal degenerate core of a WD is radiated into space through its surface. Therefore the rate of energy loss is determined to first order by the core temperature and the surface area. Higher mass WDs cool more slowly at a given temperature because their radii are smaller. In reality, cooling rates are modified by the core composition, which determines the core heat capacity, and by the composition and structure of the envelope. Several research groups have published detailed evolutionary models that provide cooling curves for a range of remnant masses (arising from the progenitor evolution) and core and envelope compositions (see e.g. Bergeron et al. 2019 and references therein). In principle, the shape of the WDLF reflects historical star formation rates moderated by the distribution of main-sequence lifetimes and subsequent WD cooling times. Furthermore, as the age of the galaxy exceeds the combined main-sequence and cooling lifetimes of the oldest white dwarfs, the cutoff at the highest absolute magnitude (lowest luminosity) can provide a low limit to the age of the disc for comparison with determinations from

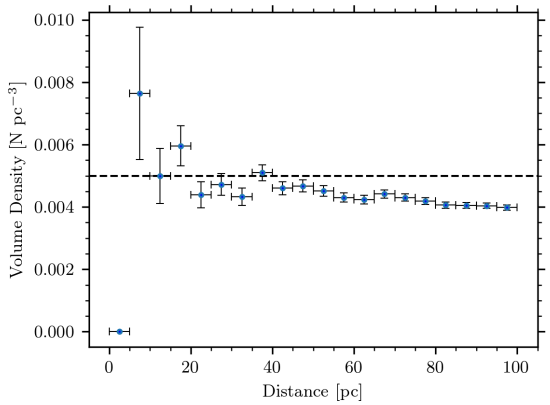

Fig. 38. Volume density of WDs as a function of distance. Values are computed for a 5 pc wide spherical shell.

other methods. The WDLF also provides insight into physical processes in WD interiors. For example, phase changes such as crystallisation release latent heat, which delays the cooling for a time. Conversely, energy loss through postulated dark matter particles (e.g. axions) might produce a detectable enhancement in cooling, if they exist.

The GCNS WD catalogue presents an opportunity to derive an WDLF without recourse to the considerably complex corrections (Lam et al. 2019) required when treating kinematically biased samples, especially those derived from reduced proper motion (Harris et al. 2006; Rowell & Hambly 2011). The GCNS sample is highly reliable and complete within a well–defined survey volume. In principle, it thus enables a straightforward derivation of the WDLF. However, the relatively low luminosity of white dwarfs compared to the apparent magnitude limit of the *Gaia* catalogue leads to some incompleteness within 100 pc. We calculated the WD volume density as a function of distance (Fig. 38). The values shown were calculated based on the number of WDs in a spherical shell with a width of 5 pc for each distance point. Within a distance of 40 pc, the WD volume density measurements show scatter from statistical number count fluctuations, but then show a clear decline by approximately 15% of the value between 40 and 100 pc, likely a consequence of the combined effect of the *Gaia* magnitude limit and a vertical decline in density in the disk.

We again employed the $1/V_{\rm Max}$ technique as detailed in Sect. 5.2 in bins of bolometric magnitude. An advantage in deriving the WDLF from *Gaia* data lies in measuring bolometric magnitudes for the fainter WDs. The white-light G passband measures a large fraction of the flux of the cooler WDs where bolometric corrections are only a weak function of effective temperature. We made the simplifying assumption that bolometric corrections ($M_{\rm bol}-M_{\rm G}$) can be taken from pure hydrogen models (we employed those of the Montreal group: Bergeron et al. 2019 and references therein) for all WDs, ignoring the effects of varying the H/He atmospheric composition and surface gravity. We interpolated amongst the model tabulated values to look up G-band bolometric corrections as a function of $(G - G_{\rm RP})$ to correct $M_{\rm G}$ to $M_{\rm bol}$. For effective temperatures $T_{\rm eff} < 4500$ K, we assumed the bolometric correction in $G = 0$ because the pure–DA model bolometric–correction–colour relationship is non–monotonic due to the effects of collisionally induced opacities in the high-pressure atmospheres. The model grids indicate that inaccuracies introduced by these simplifying assumptions are never more than a few tenths of a magnitude and are limited to the intermediately hot and very cool effective temperature ranges of the scale. Our resulting WDLF is displayed in Fig. 39, where we also show the mean $V/V_{\rm Max}$ statistic as a function of bolometric magnitude for the sample. As described in Rowell & Hambly (2011), for a sample uniformly distributed within the (generalised) survey volume, the expectation value of this statistic is $0.5 \pm 1/\sqrt{12N}$ for N stars. For the WDLF sample we find overall $V/V_{\rm Max} = 0.5050 \pm 0.0023$, with no obvious indications of systematic effects as a function of luminosity or position.

The statistical power of the GCNS sample is evident in Fig. 39. At the peak of the WDLF, nearly 1900 WDs contribute in the bin range $14.75 < M_{\rm bol} < 15.00$. There appears to be evidence of a series of features in the WDLF at high confidence: the feature around $M_{\rm bol} = 10.5$ has been noted previously (Limoges et al. 2015; Harris et al. 2006) but those at fainter levels (e.g. around $M_{\rm bol} = 14.25$) have not been so apparent. The segments between these features are linear and consistent in gradient, resulting in an apparent series of steps. The high signal-to-noise ratio and detail in this WDLF will facilitate derivation of star formation histories with inversion techniques (Rowell 2013). The peak itself appears broader than some recent determinations, and especially so with respect to simulations, although this may be the result of simplifying assumptions in such population synthesis codes as noted by Limoges et al. (2015). Furthermore, the peak appears to be slightly brighter than $M_{\rm bol} = 15$, whereas several recent determinations have reported the peak to be slightly fainter than this level. This may be an age effect, where the greater volumes studied in deep proper-motion-selected samples will net larger fractions of older thick-disc and spheroid WDs.

6. Conclusions

We have provided a well-characterised catalogue of objects within 100 pc of the Sun. In this catalogue we inferred a distance probability density function for all sources using the parallaxes and a single distance prior that takes the observational parallax cut at 8 mas and the distribution of parallax uncertainties in *Gaia* EDR3 into account. We provide all-sky maps at HEALpix level 5 of empirical magnitude limits, which we generated using all *Gaia* EDR3 entries with a G magnitude and parallax measurement. We base our magnitude limit estimator on the G magnitude distribution per HEALpix and advocate a limit between the 80th (conservative) and 90th (optimistic) percentile.

The GCNS catalogue has an estimated 331 312 entries within 100 pc. This is an increase of an order of magnitude with respect to the most complete nearby star census prior to the *Gaia* mission. A comparison with *Gaia* DR2 shows that the last release contained more contamination than *Gaia* EDR3 , but also that a few percent of real objects are still not included in *Gaia* EDR3. The overall completeness of the GCNS to M8 at 100 pc is probably better than 95%. An examination of the 10 pc sample finds that we provide the first direct parallax of five stars in multiple systems.

The GCNS was used to undertake a number of investigations into local populations, structures, and distributions. We list this below.

– We computed the luminosity function from the brightest main-sequence stars ($M_G = 2$), including part of giant stars, to the late-L brown dwarfs ($M_G = 20.5$). We found an overall density of 0.081 ± 0.003 (main sequence) stars pc^{-3}. The high signal-to-noise ratio of the luminosity function indicates features such as the Jao gap (Jao et al. 2018) and the drop in object counts at the stellar to substellar boundary (Bardalez Gagliuffi et al. 2019).

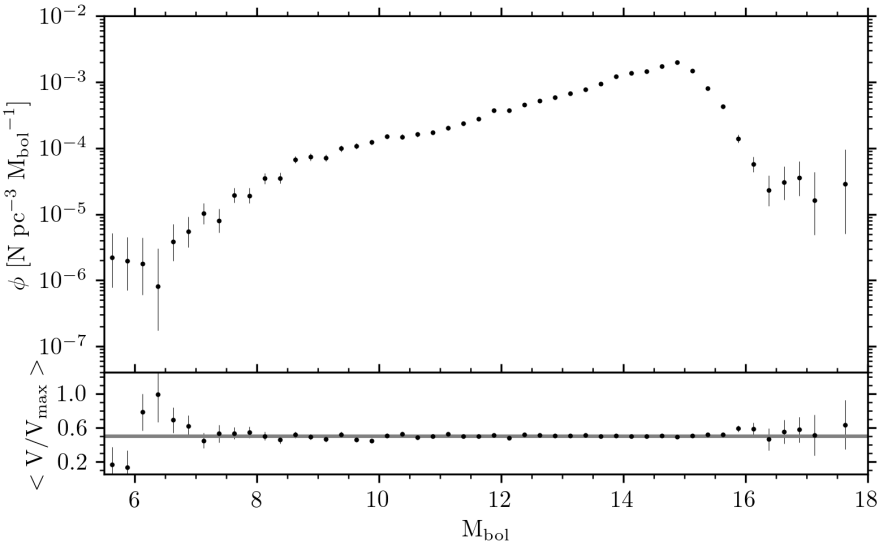

Fig. 39. *Upper panel*: WDLF for the 100 pc sample. The bin width is 0.25 $M_{\rm bol}$ and the confidence intervals are Poisson uncertainties (Gehrels 1986). The structure and features in the WDLF are statistically significant for all but the first and last few bins. *Lower panel*: $V/V_{\rm Max}$ statistic for the WDLF sample plotted in the upper panel. The expectation value for the statistic is 0.5 for a uniform sample within the survey volume (see the main text for further details).

- We explored the kinematical plane for the GCNS stars that have a *Gaia* radial velocity (74 431 stars). We show that even in the local sample, the kinematical plane shows substructures in the disc that are associated with several streams and superclusters, such as Sirius and Hercules, and in the halo, where we identified 12 stars from the *Gaia* Enceladus.
- We provide orbits for the sample. As expected, most of the stars have circular in-plane orbits similar to the Sun. However, the solar neighbourhood is also visited by several tens of stars with eccentric orbits that come from the Galactic central regions, as well as stars coming from external regions, for example the Enceladus objects.
- We briefly investigated the value for the solar motion, proposed a revision of the $V_\odot$ value to 7 km s^{-1}, and discussed the vertex deviation.
- We find 2879 new UCD candidates compared to the *Gaia* DR2, but we also note that the very nearby binary brown dwarf Luhman 16 AB does not have a five-parameter solution in the *Gaia* EDR3.
- We provided a revised catalogue of 16 556 high-probability resolved binary candidates. We confirmed the absence of bimodality in the physical projected separations distribution, placing previous DR2-based results on more solid ground. We refined the wide-binary fraction statistics as a function of spectral type, quantifying the decline in $f_{\rm WB}$ for later (K and M) spectral types.
- We re-examined the Hyades cluster and produced a list of candidate members using a procedure that did not use the GCNS selection criteria. We found only one candidate that would not have have made the GCNS.
- Using a random forest algorithm, we identified 21 848 sources with a high probability of being a WD, 2553 of which are new WD candidates.
- We derived a white dwarf luminosity function of unprecedented statistical power. Several features are clearly present that appear as a series of steps in the function. These may be indicative of variations in the historical star formation rate in the 100 pc volume and can be examined further by direction inversion techniques or comparison with population synthesis calculations.

In these investigations we have illustrated different ways of using the GCNS: the direct use of the astrometric parameters (Sects. 5.1 and 5.5), the use of derived distance PDFs (Sect. 5.2), and derived quantities (Sect. 5.3). We indicated other quality cuts that can be made to clean the catalogue using photometric flags (Sect. 5.8) and indicators of binarity (Sect. 5.2). Finally, we have shown that even though we know that the catalogue volume is incomplete, useful conclusions and constraints can be drawn (Sect. 5.4).

We expect the next releases of the *Gaia* mission to improve the GCNS in particular with the inclusion of unresolved companions and with the application of non-single star solutions in the *Gaia* processing chain where the current single-star solution will often result in erroneous astrometric parameters. In addition, the *Gaia* DR3, due to be released in 2021, will provide astrophysical parameters for nearly all the stellar sources in the *Gaia* Catalogue of Nearby Stars.

Acknowledgements. We thank the anonymous referee for comments and suggestions that improved this article. This work presents results from the European Space Agency (ESA) space mission *Gaia*. *Gaia* data are being processed by the *Gaia* Data Processing and Analysis Consortium (DPAC). Funding for the DPAC is provided by national institutions, in particular the institutions participating in the *Gaia* MultiLateral Agreement (MLA). The *Gaia* mission website is https://www.cosmos.esa.int/gaia. The *Gaia* archive website is https://archives.esac.esa.int/gaia. The *Gaia* mission and data processing have financially been supported by, in alphabetical order by country: the Algerian Centre de Recherche en Astronomie, Astrophysique et Géophysique of Bouzareah Observatory; the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (FWF) Hertha Firnberg Programme through grants T359, P20046, and P23737; the BELgian federal Science Policy Office (BELSPO) through various PROgramme de Développement d'Expériences scientifiques (PRODEX) grants and the Polish Academy of Sciences – Fonds Wetenschappelijk Onderzoek through grant VS.091.16N, and the Fonds de la Recherche Scientifique (FNRS); the Brazil-France exchange programmes Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Coordenação de Aperfeicoamento de Pessoal de Nível Superior (CAPES) – Comité Français d'Evaluation de la Coopération Universitaire et Scientifique avec le Brésil (COFECUB); the National Science Foundation of China (NSFC) through grants 11573054 and 11703065 and the China Scholarship Council through grant 201806040200; the Tenure Track Pilot Programme of the Croatian Science Foundation and the École Polytechnique Fédérale de Lausanne and the project TTP-2018-07-1171 "Mining the Variable Sky", with the funds of the Croatian-Swiss Research Programme; the Czech-Republic Ministry of Education, Youth, and Sports through grant LG 15010 and INTER-EXCELLENCE grant LTAUSA18093, and the Czech Space Office through ESA PECS contract 98058; the Danish Ministry of Science; the Estonian Ministry of Education and Research through grant IUT40-1; the European Commission's Sixth Framework Programme through the European Leadership in Space Astrometry (ELSA) Marie Curie Research Training Network (MRTN-CT-2006-033481), through Marie Curie project PIOF-GA-2009-255267 (Space AsteroSeismology & RR Lyrae stars, SAS-RRL), and through a Marie

Curie Transfer-of-Knowledge (ToK) fellowship (MTKD-CT-2004-014188); the European Commission's Seventh Framework Programme through grant FP7-606740 (FP7-SPACE-2013-1) for the *Gaia* European Network for Improved data User Services (GENIUS) and through grant 264895 for the *Gaia* Research for European Astronomy Training (GREAT-ITN) network; the European Research Council (ERC) through grants 320360 and 647208 and through the European Union's Horizon 2020 research and innovation and excellent science programmes through Marie Skłodowska-Curie grant 745617 as well as grants 670519 (Mixing and Angular Momentum tranSport of massIvE stars – MAMSIE), 687378 (Small Bodies: Near and Far), 682115 (Using the Magellanic Clouds to Understand the Interaction of Galaxies), and 695099 (A sub-percent distance scale from binaries and Cepheids – CepBin); the European Science Foundation (ESF), in the framework of the *Gaia* Research for European Astronomy Training Research Network Programme (GREAT-ESF); the European Space Agency (ESA) in the framework of the *Gaia* project, through the Plan for European Cooperating States (PECS) programme through grants for Slovenia, through contracts C98090 and 4000106398/12/NL/KML for Hungary, and through contract 4000115263/15/NL/IB for Germany; the Academy of Finland and the Magnus Ehrnrooth Foundation; the French Centre National d'Etudes Spatiales (CNES), the Agence Nationale de la Recherche (ANR) through grant ANR-10-IDEX-0001-02 for the "Investissements d'avenir" programme, through grant ANR-15-CE31-0007 for project "Modelling the Milky Way in the *Gaia* era" (MOD4Gaia), through grant ANR-14-CE33-0014-01 for project "The Milky Way disc formation in the *Gaia* era" (ARCHEOGAL), and through grant ANR-15-CE31-0012-01 for project "Unlocking the potential of Cepheids as primary distance calibrators" (UnlockCepheids), the Centre National de la Recherche Scientifique (CNRS) and its SNO *Gaia* of the Institut des Sciences de l'Univers (INSU), the "Action Fédératrice *Gaia*" of the Observatoire de Paris, the Région de Franche-Comté, and the Programme National de Gravitation, Références, Astronomie, et Métrologie (GRAM) of CNRS/INSU with the Institut National Polytechnique (INP) and the Institut National de Physique nucléaire et de Physique des Particules (IN2P3) co-funded by CNES; the German Aerospace Agency (Deutsches Zentrum für Luft- und Raumfahrt e.V., DLR) through grants 50QG0501, 50QG0601, 50QG0602, 50QG0701, 50QG0901, 50QG1001, 50QG1101, 50QG1401, 50QG1402, 50QG1403, 50QG1404, and 50QG1904 and the Centre for Information Services and High Performance Computing (ZIH) at the Technische Universität (TU) Dresden for generous allocations of computer time; the Hungarian Academy of Sciences through the Lendület Programme grants LP2014-17 and LP2018-7 and through the Premium Postdoctoral Research Programme (L. Molnár), and the Hungarian National Research, Development, and Innovation Office (NKFIH) through grant KH_18-130405; the Science Foundation Ireland (SFI) through a Royal Society - SFI University Research Fellowship (M. Fraser); the Israel Science Foundation (ISF) through grant 848/16; the Agenzia Spaziale Italiana (ASI) through contracts I/037/08/0, I/058/10/0, 2014-025-R.0, 2014-025-R.1.2015, and 2018-24-HH.0 to the Italian Istituto Nazionale di Astrofisica (INAF), contract 2014-049-R.0/1/2 to INAF for the Space Science Data Centre (SSDC, formerly known as the ASI Science Data Center, ASDC), contracts I/008/10/0, 2013/030/I.0, 2013-030-I.0.1-2015, and 2016-17-I.0 to the Aerospace Logistics Technology Engineering Company (ALTEC S.p.A.), INAF, and the Italian Ministry of Education, University, and Research (Ministero dell'Istruzione, dell'Università e della Ricerca) through the Premiale project "MIning The Cosmos Big Data and Innovative Italian Technology for Frontier Astrophysics and Cosmology" (MITiC); the Netherlands Organisation for Scientific Research (NWO) through grant NWO-M-614.061.414, through a VICI grant (A. Helmi), and through a Spinoza prize (A. Helmi), and the Netherlands Research School for Astronomy (NOVA); the Polish National Science Centre through HARMONIA grant 2018/06/M/ST9/00311, DAINA grant 2017/27/L/ST9/03221, and PRELUDIUM grant 2017/25/N/ST9/01253, and the Ministry of Science and Higher Education (MNiSW) through grant DIR/WK/2018/12; the Portugese Fundação para a Ciência e a Tecnologia (FCT) through grants SFRH/BPD/74697/2010 and SFRH/BD/128840/2017 and the Strategic Programme UID/FIS/00099/2019 for CENTRA; the Slovenian Research Agency through grant P1-0188; the Spanish Ministry of Economy (MINECO/FEDER, UE) through grants ESP2016-80079-C2-1-R, ESP2016-80079-C2-2-R, RTI2018-095076-B-C21, RTI2018-095076-B-C22, BES-2016-078499, and BES-2017-083126 and the Juan de la Cierva formación 2015 grant FJCI-2015-2671, the Spanish Ministry of Education, Culture, and Sports through grant FPU16/03827, the Spanish Ministry of Science and Innovation (MICINN) through grant AYA2017-89841P for project "Estudio de las propiedades de los fósiles estelares en el entorno del Grupo Local" and through grant TIN2015-65316-P for project "Computación de Altas Prestaciones VII", the Severo Ochoa Centre of Excellence Programme of the Spanish Government through grant SEV2015-0493, the Institute of Cosmos Sciences University of Barcelona (ICCUB, Unidad de Excelencia "María de Maeztu") through grants MDM-2014-0369 and CEX2019-000918-M, the University of Barcelona's official doctoral programme for the development of an R+D+i project through an Ajuts de Personal Investigador en Formació (APIF) grant, the Spanish Virtual Observatory through project AyA2017-84089, the Galician Regional Government, Xunta de Galicia, through grants ED431B-2018/42 and ED481A-2019/155, support received from the Centro de Investigación en Tecnologías de la Información y las Comunicaciones (CITIC) funded by the Xunta de Galicia, the Xunta de Galicia and the Centros Singulares de Investigación de Galicia for the period 2016-2019 through CITIC, the European Union through the European Regional Development Fund (ERDF) / Fondo Europeo de Desenvolvemento Rexional (FEDER) for the Galicia 2014-2020 Programme through grant ED431G-2019/01, the Red Española de Supercomputación (RES) computer resources at MareNostrum, the Barcelona Supercomputing Centre – Centro Nacional de Supercomputación (BSC-CNS) through activities AECT-2016-1-0006, AECT-2016-2-0013, AECT-2016-3-0011, and AECT-2017-1-0020, the Departament d'Innovació, Universitats i Empresa de la Generalitat de Catalunya through grant 2014-SGR-1051 for project "Models de Programació i Entorns d'Execució Parallels" (MPEXPAR), and Ramon y Cajal Fellowship RYC2018-025968-I; the Swedish National Space Agency (SNSA/Rymdstyrelsen); the Swiss State Secretariat for Education, Research, and Innovation through the Mesures d'Accompagnement, the Swiss Activités Nationales Complémentaires, and the Swiss National Science Foundation; the United Kingdom Particle Physics and Astronomy Research Council (PPARC), the United Kingdom Science and Technology Facilities Council (STFC), and the United Kingdom Space Agency (UKSA) through the following grants to the University of Bristol, the University of Cambridge, the University of Edinburgh, the University of Leicester, the Mullard Space Sciences Laboratory of University College London, and the United Kingdom Rutherford Appleton Laboratory (RAL): PP/D006511/1, PP/D006546/1, PP/D006570/1, ST/I000852/1, ST/J005045/1, ST/K00056X/1, ST/K000209/1, ST/K000756/1, ST/L006561/1, ST/N000595/1, ST/N000641/1, ST/N000978/1, ST/N001117/1, ST/S000089/1, ST/S000976/1, ST/S001123/1, ST/S001948/1, ST/S002103/1, and ST/V000969/1. The *Gaia* project, data processing and this contribution have made use of: the Set of Identifications, Measurements, and Bibliography for Astronomical Data (SIMBAD, Wenger et al. 2000), the "Aladin sky atlas" (Bonnarel et al. 2000; Boch & Fernique 2014), and the VizieR catalogue access tool (Ochsenbein et al. 2000), all operated at the Centre de Données astronomiques de Strasbourg (CDS); the National Aeronautics and Space Administration (NASA) Astrophysics Data System (ADS); the software products TOPCAT, and STILTS (Taylor 2005, 2006); Astropy, a community-developed core Python package for Astronomy (Astropy Collaboration 2018); data products from the Two Micron All Sky Survey (2MASS, Skrutskie et al. 2006), which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center (IPAC) / California Institute of Technology, funded by the National Aeronautics and Space Administration (NASA) and the National Science Foundation (NSF) of the USA; the first data release of the Pan-STARRS survey (Chambers et al. 2016) The Pan-STARRS1 Surveys (PS1) and the PS1 public science archive have been made possible through contributions by the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, the Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration (NASA) through grant NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation through grant AST-1238877, the University of Maryland, Eotvos Lorand University (ELTE), the Los Alamos National Laboratory, and the Gordon and Betty Moore Foundation; data products from the Wide-field Infrared Survey Explorer (WISE), which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, and NEOWISE, which is a project of the Jet Propulsion Laboratory/California Institute of Technology. WISE and NEOWISE are funded by the National Aeronautics and Space Administration (NASA); the fifth data release of the Radial Velocity Experiment (RAVE DR5, Kunder et al. 2017). Funding for RAVE has been provided by the Australian Astronomical Observatory, the Leibniz-Institut für Astrophysik Potsdam (AIP), the Australian National University, the Australian Research Council, the French National Research Agency, the German Research Foundation (SPP 1177 and SFB 881), the European Research Council (ERC-StG 240271 Galactica), the Istituto Nazionale di Astrofisica at Padova, The Johns Hopkins University, the National Science Foundation of the USA (AST-0908326), the W. M. Keck foundation, the Macquarie University, the Netherlands Research School for Astronomy, the Natural Sciences and Engineering Research Council of Canada, the Slovenian Research Agency, the Swiss National Science Foundation, the Science & Technology Facilities Council of the UK, Opticon, Strasbourg Observatory, and the Universities of Groningen, Heidelberg, and Sydney. The RAVE website is at https://www.rave-survey.org/; the thirteenth release of the Sloan Digital Sky Survey (SDSS DR13, Albareti et al. 2017). Funding for SDSS-IV has been provided by the Alfred P. Sloan Foundation, the United States Department of Energy Office of Science, and the

Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is https://www.sdss.org/. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University; the second release of the SkyMapper catalogue (SkyMapper DR2, Onken et al. 2019, Digital Object Identifier 10.25914/5ce60d31ce759). The national facility capability for SkyMapper has been funded through grant LE130100104 from the Australian Research Council (ARC) Linkage Infrastructure, Equipment, and Facilities (LIEF) programme, awarded to the University of Sydney, the Australian National University, Swinburne University of Technology, the University of Queensland, the University of Western Australia, the University of Melbourne, Curtin University of Technology, Monash University, and the Australian Astronomical Observatory. SkyMapper is owned and operated by The Australian National University's Research School of Astronomy and Astrophysics. The survey data were processed and provided by the SkyMapper Team at the Australian National University. The SkyMapper node of the All-Sky Virtual Observatory (ASVO) is hosted at the National Computational Infrastructure (NCI). Development and support the SkyMapper node of the ASVO has been funded in part by Astronomy Australia Limited (AAL) and the Australian Government through the Commonwealth's Education Investment Fund (EIF) and National Collaborative Research Infrastructure Strategy (NCRIS), particularly the National eResearch Collaboration Tools and Resources (NeCTAR) and the Australian National Data Service Projects (ANDS).

References

Albareti, F. D., Allende Prieto, C., Almeida, A., et al. 2017, ApJS, 233, 25
Andrews, J. J., Chanamé, J., & Agüeros, M. A. 2017, MNRAS, 472, 675
Antoja, T., Valenzuela, O., Pichardo, B., et al. 2009, ApJ, 700, L78
Antoja, T., Figueras, F., Torra, J., Valenzuela, O., & Pichardo, B. 2010, The Origin of Stellar Moving Groups, (Difusion Centro de Publicacion y Publica), 4, 13
Antoja, T., Helmi, A., Bienayme, O., et al. 2012, MNRAS, 426, L1
Antoja, T., Helmi, A., Dehnen, W., et al. 2014, A&A, 563, A60
Arenou, F. 2010, GAIA-C2-SP-OPM-FA-054
Arenou, F. 2011, AIP Conf. Ser., 1346, 107
Astropy Collaboration (Price-Whelan, A. M., et al.) 2018, AJ, 156, 123
Baraffe, I., & Chabrier, G. 2018, A&A, 619, A177
Baraffe, I., Homeier, D., Allard, F., & Chabrier, G. 2015, A&A, 577, A42
Bardalez Gagliuffi, D. C., Burgasser, A. J., Schmidt, S. J., et al. 2019, ApJ, 883, 205
Belokurov, V., Penoyre, Z., Oh, S., et al. 2020, MNRAS, 496, 1922
Bergeron, P., Dufour, P., Fontaine, G., et al. 2019, ApJ, 876, 67
Bienayme, O., & Sechaud, N. 1997, A&A, 323, 781
Bienaymé, O., Robin, A. C., & Famaey, B. 2015, A&A, 581, A123
Bienaymé, O., Leca, J., & Robin, A. C. 2018, A&A, 620, A103
Biller, B. A., Close, L. M., Masciadri, E., et al. 2007, ApJS, 173, 143
Boch, T., & Fernique, P. 2014, ASP Conf. Ser., 485, 277
Bonnarel, F., Fernique, P., Bienaymé, O., et al. 2000, A&AS, 143, 33
Boubert, D., & Everall, A. 2020, MNRAS, 497, 4246
Bovy, J., Allende Prieto, C., Beers, T. C., et al. 2012, ApJ, 759, 131
Brandeker, A., & Cataldi, G. 2019, A&A, 621, A86
Breiman, L. 2001, Mach. Learn., 45, 5-32
Breiman, L. 2002, unpublished technical note
Burgasser, A. J., Kirkpatrick, J. D., Reid, I. N., et al. 2003, ApJ, 586, 512
Caballero, J. A. 2010, A&A, 514, A98
Carney, B. W., Laird, J. B., Latham, D. W., & Aguilar, L. A. 1996, AJ, 112, 668
Carpenter, B., Gelman, A., Hoffman, M., et al. 2017, J. Stat. Softw. Art,, 76, 1
Carrasco, J. M., Evans, D. W., Montegriffo, P., et al. 2016, A&A, 595, A7
Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, ArXiv e-prints [arXiv:1612.05560]
Chereul, E., Creze, M., & Bienayme, O. 1997, ESA SP, 402, 545
Chumak, Y. O., Rastorguev, A. S., & Aarseth, S. J. 2005, Astron. Lett., 31, 308
Clemens, J. C., Reid, I. N., Gizis, J. E., & O'Brien, M. S. 1998, ApJ, 496, 352
Creze, M., & Mennessier, M. O. 1973, A&A, 27, 281
Cruz, K. L., Reid, I. N., Kirkpatrick, J. D., et al. 2007, AJ, 133, 439
Cutri, R. M., Wright, E. L., Conrow, T., et al. 2013, Explanatory Supplement to the AllWISE Data Release Products
Czekaj, M. A., Robin, A. C., Figueras, F., Luri, X., & Haywood, M. 2014, A&A, 564, A102
Deacon, N. R., Kraus, A. L., Mann, A. W., et al. 2016, MNRAS, 455, 4212
de Bruijne J. H. J. 2012, Ap&SS, 341, 31
de Bruijne, J. H. J., Allen, M., Azaz, S., et al. 2015, A&A, 576, A74
Dehnen, W. 1998, AJ, 115, 2384
Dehnen, W., & Binney, J. J. 1998, MNRAS, 298, 387
Desidera, S., & Barbieri, M. 2007, A&A, 462, 345
Dittmann, J. A., Irwin, J. M., Charbonneau, D., & Berta-Thompson, Z. K. 2014, ApJ, 784, 156
Dobbie, P. S., & Warren, S. J. 2020, Open J. Astrophys., 3, 5
Duquennoy, A., & Mayor, M. 1991, A&A, 500, 337
Eggen, O. J. 1958, The Observatory, 78, 21
Eggen, O. J. 1971, PASP, 83, 251
Eilers, A.-C., Hogg, D. W., Rix, H.-W., & Ness, M. K. 2019, ApJ, 871, 120
Einasto, J. 1979, IAU Symp., 84, 451
Eisenhardt, P. R. M., Marocco, F., Fowler, J. W., et al. 2020, ApJS, 247, 69
El-Badry, K., & Rix, H.-W. 2018, MNRAS, 480, 4884
ESA 1997, ESA SP, 1200, The HIPPARCOS and TYCHO catalogues. Astrometric and photometric star catalogues derived from the ESA HIPPARCOS Space Astrometry Mission
Evans, D. W., Riello, M., De Angeli, F., et al. 2018, A&A, 616, A4
Faherty, J. K., Bochanski, J. J., Gagné, J., et al. 2018, ApJ, 863, 91
Famaey, B., Siebert, A., & Jorissen, A. 2008, A&A, 483, 453
Felten, J. E. 1976, ApJ, 207, 700
Fernandez-Trincado, J. 2017, PhD thesis, http://theses.fr/s108979, University Bourgogne-Franche-Comté, France
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306
Friel, E., Cayrel de Strobel, G., Chmielewski, Y., et al. 1993, A&A, 274, 825
Gaia Collaboration (Brown, A. G. A., et al.) 2016, A&A, 595, A2
Gaia Collaboration (Babusiaux, C., et al.) 2018a, A&A, 616, A10
Gaia Collaboration (Brown, A. G. A., et al.) 2018b, A&A, 616, A1
Gaia Collaboration (Helmi, A., et al.) 2018c, A&A, 616, A12
Gaia Collaboration (Katz, D., et al.) 2018d, A&A, 616, A11
Gaia Collaboration (Brown, A. G. A., et al.) 2021, A&A, 649, A1 (*Gaia* EDR3 SI)
Gehrels, N. 1986, ApJ, 303, 336
Gentile Fusillo, N. P., Tremblay, P.-E., Gänsicke, B. T., et al. 2019, MNRAS, 482, 4570
Gini, C. 1912, Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche. [Fasc. I.], Studi economico-giuridici pubblicati per cura della facoltà di Giurisprudenza della R. Università di Cagliari (Tipogr. di P. Cuppini)
Gizis, J. E., & Reid, I. N. 1999, AJ, 117, 508
Gliese, W. 1957, Astronomisches Rechen-Institut Heidelberg Mitteilungen Serie A, 8, 1
Gliese, W., & Jahreiß, H. 1991, Preliminary Version of the Third Catalogue of Nearby Stars, On: The Astronomical Data Center CD-ROM: Selected Astronomical Catalogs
Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
Harris, H. C., Munn, J. A., Kilic, M., et al. 2006, AJ, 131, 571
Hartman, Z. D., & Lépine, S. 2020, ApJS, 247, 66
Hastie, T., & Stuetzle, W. 1989, J. Am. Stat. Assoc., 84, 502
Helmi, A., Babusiaux, C., Koppelman, H. H., et al. 2018, Nature, 563, 85
Henry, T. J., Franz, O. G., Wasserman, L. H., et al. 1999, ApJ, 512, 864
Henry, T. J., Jao, W.-C., Winters, J. G., et al. 2018, AJ, 155, 265
Hunt, J. A. S., Hong, J., Bovy, J., Kawata, D., & Grand, R. J. J. 2018, MNRAS, 481, 3794
Jaehnig, K., Somers, G., & Stassun, K. G. 2019, ApJ, 879, 39
Jao, W.-C., & Feiden, G. A. 2020, AJ, 160, 102
Jao, W.-C., Henry, T. J., Gies, D. R., & Hambly, N. C. 2018, ApJ, 861, L11
Jiménez-Esteban, F. M., Torres, S., Rebassa-Mansergas, A., et al. 2018, MNRAS, 480, 4505
Jiménez-Esteban, F. M., Solano, E., & Rodrigo, C. 2019, AJ, 157, 78
Johnson, J. A., Butler, R. P., Marcy, G. W., et al. 2007, ApJ, 670, 833
Just, A., Fuchs, B., Jahreiß, H., et al. 2015, MNRAS, 451, 149
Kane, S. R., Dalba, P. A., Li, Z., et al. 2019, AJ, 157, 252
Karim, M. T., & Mamajek, E. E. 2017, MNRAS, 465, 472

Kharchenko, N. V., Berczik, P., Petrov, M. I., et al. 2009, A&A, 495, 807
Kirkpatrick, J. D., Henry, T. J., & Irwin, M. J. 1997, AJ, 113, 1421
Kirkpatrick, J. D., Gelino, C. R., Cushing, M. C., et al. 2012, ApJ, 753, 156
Kirkpatrick, J. D., Martin, E. C., Smart, R. L., et al. 2019, ApJS, 240, 19
Kraus, A. L., Ireland, M. J., Huber, D., Mann, A. W., & Dupuy, T. J. 2016, AJ, 152, 8
Kroupa, P., Tout, C. A., & Gilmore, G. 1990, MNRAS, 244, 76
Kunder, A., Kordopatis, G., Steinmetz, M., et al. 2017, AJ, 153, 75
Lagarde, N., Robin, A. C., Reylé, C., & Nasello, G. 2017, A&A, 601, A27
Laithwaite, R. C., & Warren, S. J. 2020, MNRAS, 499, 2587
Lam, M. C., Hambly, N. C., Rowell, N., et al. 2019, MNRAS, 482, 715
Latham, D. W., Mazeh, T., Davis, R. J., Stefanik, R. P., & Abt, H. A. 1991, AJ, 101, 625
Limoges, M. M., Bergeron, P., & Lépine, S. 2015, ApJS, 219, 19
Lindegren, L., Hernández, J., Bombrun, A., et al. 2018, A&A, 616, A2
Lindegren, L., Bastian, U., Biermann, M., et al. 2021a, A&A, 649, A4 (*Gaia* EDR3 SI)
Lindegren, L., Klioner, S. A., Hernández, J., et al. 2021b, A&A, 649, A2 (*Gaia* EDR3 SI)
Lodieu, N., Smart, R. L., Pérez-Garrido, A., & Silvotti, R. 2019, A&A, 623, A35
Lowrance, P. J., Kirkpatrick, J. D., & Beichman, C. A. 2002, ApJ, 572, L79
Luhman, K. L. 2013, ApJ, 767, L1
Luri, X., Palmer, M., Arenou, F., et al. 2014, A&A, 566, A119
Lutz, T. E., & Kelker, D. H. 1973, PASP, 85, 573
MacDonald, J., & Gizis, J. 2018, MNRAS, 480, 1711
Marrese, P. M., Marinoni, S., Fabrizio, M., & Altavilla, G. 2019, A&A, 621, A144
Martin, E. C., Kirkpatrick, J. D., Beichman, C. A., et al. 2018, ApJ, 867, 109
Martini, P., & Osmer, P. S. 1998, AJ, 116, 2513
Mayor, M. 1970, A&A, 6, 60
Meingast, S., & Alves, J. 2019, A&A, 621, L3
Merle, T., Van der Swaelmen, M., Van Eck, S., et al. 2020, A&A, 635, A155
Michtchenko, T. A., Lépine, J. R. D., Pérez-Villegas, A., Vieira, R. S. S., & Barros, D. A. 2018, ApJ, 863, L37
Moe, M., & Di Stefano R. 2017, ApJS, 230, 15
Moe, M., Kratter, K. M., & Badenes, C. 2019, ApJ, 875, 61
Monari, G., Kawata, D., Hunt, J. A. S., & Famaey, B. 2017, MNRAS, 466, L113
Mor, R., Robin, A. C., Figueras, F., & Antoja, T. 2018, A&A, 620, A79
Mor, R., Robin, A. C., Figueras, F., Roca-Fàbrega, S., & Luri, X. 2019, A&A, 624, L1
Ochsenbein, F., Bauer, P., & Marcout, J. 2000, A&AS, 143, 23
Oelkers, R. J., Stassun, K. G., & Dhital, S. 2017, AJ, 153, 259
Oh, S., & Evans, N. W. 2020, MNRAS, 498, 1920
Oh, S., Price-Whelan, A. M., Hogg, D. W., Morton, T. D., & Spergel, D. N. 2017, AJ, 153, 257
Onken, C. A., Wolf, C., Bessell, M. S., et al. 2019, PASA, 36, e033
Perryman, M. A. C., Lindegren, L., Kovalevsky, J., et al. 1997, A&A, 323, L49
Poveda, A., Herrera, M. A., Allen, C., Cordero, G., & Lavalley, C. 1994, Rev. Mex. Astron. Astrofis., 28, 43
Pravdo, S. H., Shaklan, S. B., Wiktorowicz, S. J., et al. 2006, ApJ, 649, 389
Price-Whelan, A. M., Hogg, D. W., Rix, H.-W., et al. 2020, ApJ, 895, 2
Raghavan, D., McAlister, H. A., Henry, T. J., et al. 2010, ApJS, 190, 1
Reggiani, H., & Meléndez, J. 2018, MNRAS, 475, 3502
Reid, I. N., & Gizis, J. E. 1997, AJ, 113, 2246
Reid, I. N., Gizis, J. E., & Hawley, S. L. 2002, AJ, 124, 2721
Reid, I. N., Cruz, K. L., Laurie, S. P., et al. 2003, AJ, 125, 354
Reino, S., de Bruijne, J., Zari, E., d'Antona, F., & Ventura, P. 2018, MNRAS, 477, 3197
Reylé, C. 2018, A&A, 619, L8
Reylé, C., Delorme, P., Willott, C. J., et al. 2010, A&A, 522, A112
Riello, M., De Angeli, F., Evans, D. W., et al. 2018, A&A, 616, A3
Riello, M., De Angeli, F., Evans, D. W., et al. 2021, A&A, 649, A3 (*Gaia* EDR3 SI)
Rix, H.-W., & Bovy, J. 2013, A&ARv, 21, 61
Robin, A. C., Reylé, C., Derrière, S., & Picaud, S. 2003, A&A, 409, 523
Robin, A. C., Luri, X., Reylé, C., et al. 2012, A&A, 543, A100
Robin, A. C., Bienaymé, O., Fernández-Trincado, J. G., & Reylé, C. 2017, A&A, 605, A1
Röser, S., Schilbach, E., & Goldman, B. 2019, A&A, 621, L2
Rowell, N. 2013, MNRAS, 434, 1549
Rowell, N., & Hambly, N. C. 2011, MNRAS, 417, 93
Rybizki, J., & Drimmel, R. 2018, gdr2_completeness: GaiaDR2 data retrieval and manipulation
Rybizki, J., Demleitner, M., Bailer-Jones, C., et al. 2020, PASP, 132, 074501
Schmidt, M. 1968, ApJ, 151, 393
Scholz, R. D. 2020, A&A, 637, A45
Schönrich, R., Binney, J., & Dehnen, W. 2010, MNRAS, 403, 1829
Seabroke, G., Fabricius, C., Teyssier, D., et al. 2021, A&A, submitted (*Gaia* EDR3 SI)
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Skuljan, J., Hearnshaw, J. B., & Cottrell, P. L. 1999, MNRAS, 308, 731
Smart, R. L., Tinney, C. G., Bucciarelli, B., et al. 2013, MNRAS, 433, 2054
Smart, R. L., Marocco, F., Caballero, J. A., et al. 2017, MNRAS, 469, 401
Smart, R. L., Marocco, F., Sarro, L. M., et al. 2019, MNRAS, 485, 4423
Söderhjelm, S. 2004, ASP Conf. Ser., 318, 413
Söderhjelm, S. 2005, ESA SP, 576, 97
Sozzetti, A., Torres, G., Latham, D. W., et al. 2009, ApJ, 697, 544
Stauffer, J. R. 1984, ApJ, 280, 189
Stauffer, J., Tanner, A. M., Bryden, G., et al. 2010, PASP, 122, 885
Tachibana, Y., & Miller, A. A. 2018, PASP, 130, 128001
Tang, S.-Y., Chen, W. P., Chiang, P. S., et al. 2018, ApJ, 862, 106
Tang, S.-Y., Pang, X., Yuan, Z., et al. 2019, ApJ, 877, 12
Taylor, M. B. 2005, ASP Conf. Ser., 347, 29
Taylor, M. B. 2006, ASP Conf. Ser., 351, 666
Tinney, C. G., Reid, I. N., & Mould, J. R. 1993, ApJ, 414, 254
Tokovinin, A. 2018, ApJS, 235, 6
Torres, S., Cantero, C., Rebassa-Mansergas, A., et al. 2019, MNRAS, 485, 5573
van Altena, W. F., Lee, J. T., & Hoffleit, E. D. 1995, The general catalogue of trigonometric [stellar] parallaxes (Yale University Observatory)
van Leeuwen, F. 2007, A&A, 474, 653
Vrijmoet, E. H., Henry, T. J., Jao, W.-C., & Dieterich, S. B. 2020, AJ, 160, 215
Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, A&AS, 143, 9
Widmark, A., Leistedt, B., & Hogg, D. W. 2018, ApJ, 857, 114
Wielen, R. 1974, Highlights Astron., 3, 395
Wielen, R., Jahreiß, H., & Krüger, R. 1983, IAU Colloq. 76: Nearby Stars and the Stellar Luminosity Function, eds. A. G. D. Philip & A. R. Upgren, 163
Winters, J. G., Henry, T. J., Jao, W.-C., et al. 2019, AJ, 157, 216
Wolf, C., Onken, C. A., Luvaul, L. C., et al. 2018, PASA, 35, e010
Zari, E., Hashemi, H., Brown, A. G. A., Jardine, K., & de Zeeuw P. T. 2018, A&A, 620, A172

[1] INAF – Osservatorio Astrofisico di Torino, Via Osservatorio 20, 10025 Pino Torinese (TO), Italy
[2] Departmento de Inteligencia Artificial, UNED, c/ Juan del Rosal 16, 28040 Madrid, Spain
[3] Max Planck Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany
[4] Institut UTINAM CNRS UMR6213, Université Bourgogne Franche-Comté, OSU THETA Franche-Comté Bourgogne, Observatoire de Besançon, BP1615, 25010 Besançon Cedex, France
[5] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[6] School of Physics and Astronomy, University of Leicester, University Road, Leicester LE1 7RH, UK
[7] European Space Agency (ESA), European Space Research and Technology Centre (ESTEC), Keplerlaan 1, 2201AZ, Noordwijk, The Netherlands
[8] Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, 08028 Barcelona, Spain
[9] Centre for Astrophysics Research, University of Hertfordshire, College Lane, AL10 9AB, Hatfield, UK
[10] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
[11] RHEA for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[12] Leiden Observatory, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands
[13] INAF – Osservatorio astronomico di Padova, Vicolo Osservatorio 5, 35122 Padova, Italy
[14] Univ. Grenoble Alpes, CNRS, IPAG, 38000 Grenoble, France
[15] GEPI, Observatoire de Paris, Université PSL, CNRS, 5 Place Jules Janssen, 92190 Meudon, France
[16] Astronomisches Rechen-Institut, Zentrum für Astronomie der Universität Heidelberg, Mönchhofstr. 12-14, 69120 Heidelberg, Germany
[17] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Bd de l'Observatoire, CS 34229, 06304 Nice Cedex 4, France

[18] Department of Astronomy, University of Geneva, Chemin des Maillettes 51, 1290 Versoix, Switzerland
[19] Aurora Technology for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[20] Lohrmann Observatory, Technische Universität Dresden, Mommsenstraße 13, 01062 Dresden, Germany
[21] European Space Agency (ESA), European Space Astronomy Centre (ESAC), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[22] Lund Observatory, Department of Astronomy and Theoretical Physics, Lund University, Box 43, 22100 Lund, Sweden
[23] CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[24] Institut d'Astronomie et d'Astrophysique, Université Libre de Bruxelles CP 226, Boulevard du Triomphe, 1050 Brussels, Belgium
[25] F.R.S.-FNRS, Rue d'Egmont 5, 1000 Brussels, Belgium
[26] INAF – Osservatorio Astrofisico di Arcetri, Largo Enrico Fermi 5, 50125 Firenze, Italy
[27] Laboratoire d'astrophysique de Bordeaux, Univ. Bordeaux, CNRS, B18N, allée Geoffroy Saint-Hilaire, 33615 Pessac, France
[28] Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK
[29] University of Turin, Department of Physics, Via Pietro Giuria 1, 10125 Torino, Italy
[30] DAPCOM for Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, 08028 Barcelona, Spain
[31] Royal Observatory of Belgium, Ringlaan 3, 1180 Brussels, Belgium
[32] ALTEC S.p.a, Corso Marche, 79, 10146 Torino, Italy
[33] Department of Astronomy, University of Geneva, Chemin d'Ecogia 16, 1290 Versoix, Switzerland
[34] Sednai Sàrl, Geneva, Switzerland
[35] Gaia DPAC Project Office, ESAC, Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[36] Telespazio Vega UK Ltd for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[37] SYRTE, Observatoire de Paris, Université PSL, CNRS, Sorbonne Université, LNE, 61 avenue de l'Observatoire 75014 Paris, France
[38] National Observatory of Athens, I. Metaxa and Vas. Pavlou, Palaia Penteli, 15236 Athens, Greece
[39] IMCCE, Observatoire de Paris, Université PSL, CNRS, Sorbonne Université, Univ. Lille, 77 av. Denfert-Rochereau, 75014 Paris, France
[40] INAF – Osservatorio Astrofisico di Catania, Via S. Sofia 78, 95123 Catania, Italy
[41] Serco Gestión de Negocios for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[42] INAF – Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Piero Gobetti 93/3, 40129 Bologna, Italy
[43] Institut d'Astrophysique et de Géophysique, Université de Liège, 19c, Allée du 6 Août, 4000 Liège, Belgium
[44] CRAAG – Centre de Recherche en Astronomie, Astrophysique et Géophysique, Route de l'Observatoire Bp 63 Bouzareah 16340 Algiers, Algeria
[45] ATG Europe for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[46] ETSE Telecomunicación, Universidade de Vigo, Campus Lagoas-Marcosende, 36310 Vigo, Galicia, Spain
[47] Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550, 11 rue de l'Université, 67000 Strasbourg, France
[48] Kavli Institute for Cosmology Cambridge, Institute of Astronomy, Madingley Road, Cambridge, CB3 0HA, USA
[49] Department of Astrophysics, Astronomy and Mechanics, National and Kapodistrian University of Athens, Panepistimiopolis, Zografos, 15783 Athens, Greece
[50] Observational Astrophysics, Division of Astronomy and Space Physics, Department of Physics and Astronomy, Uppsala University, Box 516, 751 20 Uppsala, Sweden
[51] Leibniz Institute for Astrophysics Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany
[52] CENTRA, Faculdade de Ciências, Universidade de Lisboa, Edif. C8, Campo Grande, 1749-016 Lisboa, Portugal
[53] Department of Informatics, Donald Bren School of Information and Computer Sciences, University of California, 5019 Donald Bren Hall, 92697-3440 CA Irvine, USA
[54] Dipartimento di Fisica e Astronomia "Ettore Majorana", Università di Catania, Via S. Sofia 64, 95123 Catania, Italy
[55] CITIC, Department of Nautical Sciences and Marine Engineering, University of A Coruña, Campus de Elviña s/n, 15071, A Coruña, Spain
[56] INAF – Osservatorio Astronomico di Roma, Via Frascati 33, 00078 Monte Porzio Catone (Roma), Italy
[57] Space Science Data Center – ASI, Via del Politecnico SNC, 00133 Roma, Italy
[58] Department of Physics, University of Helsinki, PO Box 64, 00014 Helsinki, Finland
[59] Finnish Geospatial Research Institute FGI, Geodeetinrinne 2, 02430 Masala, Finland
[60] STFC, Rutherford Appleton Laboratory, Harwell, Didcot, OX11 0QX, UK
[61] HE Space Operations BV for European Space Agency (ESA), Keplerlaan 1, 2201AZ, Noordwijk, The Netherlands
[62] Applied Physics Department, Universidade de Vigo, 36310 Vigo, Spain
[63] Thales Services for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[64] Instituut voor Sterrenkunde, KU Leuven, Celestijnenlaan 200D, 3001 Leuven, Belgium
[65] Department of Astrophysics/IMAPP, Radboud University, PO Box 9010, 6500 GL Nijmegen, The Netherlands
[66] CITIC – Department of Computer Science and Information Technologies, University of A Coruña, Campus de Elviña s/n, 15071, A Coruña, Spain
[67] Barcelona Supercomputing Center (BSC) – Centro Nacional de Supercomputación, c/ Jordi Girona 29, Ed. Nexus II, 08034 Barcelona, Spain
[68] University of Vienna, Department of Astrophysics, Türkenschanzstraße 17, A1180 Vienna, Austria
[69] European Southern Observatory, Karl-Schwarzschild-Str. 2, 85748 Garching, Germany
[70] Kapteyn Astronomical Institute, University of Groningen, Landleven 12, 9747 AD Groningen, The Netherlands
[71] Center for Research and Exploration in Space Science and Technology, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore MD, USA
[72] GSFC – Goddard Space Flight Center, Code 698, 8800 Greenbelt Rd, 20771 MD Greenbelt, USA
[73] EURIX S.r.l., Corso Vittorio Emanuele II 61, 10128 Torino, Italy
[74] Harvard-Smithsonian Center for Astrophysics, 60 Garden St., MS 15, Cambridge, MA 02138, USA
[75] HE Space Operations BV for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[76] CAUP – Centro de Astrofisica da Universidade do Porto, Rua das Estrelas, Porto, Portugal
[77] SISSA – Scuola Internazionale Superiore di Studi Avanzati, Via Bonomea 265, 34136 Trieste, Italy
[78] Telespazio for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[79] University of Turin, Department of Computer Sciences, Corso Svizzera 185, 10149 Torino, Italy

[80] Departmento de Matemática Aplicada y Ciencias de la Computación, Univ. de Cantabria, ETS Ingenieros de Caminos, Canales y Puertos, Avda. de los Castros s/n, 39005 Santander, Spain
[81] Centro de Astronomía – CITEVA, Universidad de Antofagasta, Avenida Angamos 601, Antofagasta 1270300, Chile
[82] Vera C Rubin Observatory, 950 N. Cherry Avenue, Tucson, AZ 85719, USA
[83] University of Antwerp, Onderzoeksgroep Toegepaste Wiskunde, Middelheimlaan 1, 2020 Antwerp, Belgium
[84] INAF – Osservatorio Astronomico d'Abruzzo, Via Mentore Maggini, 64100 Teramo, Italy
[85] Instituto de Astronomia, Geofìsica e Ciências Atmosféricas, Universidade de São Paulo, Rua do Matão, 1226, Cidade Universitaria, 05508-900 São Paulo, SP, Brazil
[86] Mésocentre de calcul de Franche-Comté, Université de Franche-Comté, 16 route de Gray, 25030 Besançon Cedex, France
[87] SRON, Netherlands Institute for Space Research, Sorbonnelaan 2, 3584CA, Utrecht, The Netherlands
[88] Theoretical Astrophysics, Division of Astronomy and Space Physics, Department of Physics and Astronomy, Uppsala University, Box 516, 751 20 Uppsala, Sweden
[89] ATOS for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[90] School of Physics and Astronomy, Tel Aviv University, Tel Aviv 6997801, Israel
[91] Astrophysics Research Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast BT7 1NN, UK
[92] Centre de Données Astronomique de Strasbourg, Strasbourg, France
[93] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Géoazur, Bd de l'Observatoire, CS 34229, 06304 Nice Cedex 4, France
[94] Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Straße 1, 85748 Garching, Germany
[95] APAVE SUDEUROPE SAS for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[96] Área de Lenguajes y Sistemas Informáticos, Universidad Pablo de Olavide, Ctra. de Utrera, km 1. 41013, Sevilla, Spain
[97] Onboard Space Systems, Luleå University of Technology, Box 848, 981 28 Kiruna, Sweden
[98] TRUMPF Photonic Components GmbH, Lise-Meitner-Straße 13, 89081 Ulm, Germany
[99] IAC – Instituto de Astrofisica de Canarias, Via Láctea s/n, 38200 La Laguna S.C., Tenerife, Spain
[100] Department of Astrophysics, University of La Laguna, Via Láctea s/n, 38200 La Laguna S.C., Tenerife, Spain
[101] Laboratoire Univers et Particules de Montpellier, CNRS Université Montpellier, Place Eugène Bataillon, CC72, 34095 Montpellier Cedex 05, France
[102] LESIA, Observatoire de Paris, Université PSL, CNRS, Sorbonne Université, Université de Paris, 5 Place Jules Janssen, 92190 Meudon, France
[103] Villanova University, Department of Astrophysics and Planetary Science, 800 E Lancaster Avenue, Villanova PA 19085, USA
[104] Astronomical Observatory, University of Warsaw, Al. Ujazdowskie 4, 00-478 Warszawa, Poland
[105] Laboratoire d'astrophysique de Bordeaux, Univ. Bordeaux, CNRS, B18N, allée Geoffroy Saint-Hilaire, 33615 Pessac, France
[106] Université Rennes, CNRS, IPR (Institut de Physique de Rennes) - UMR 6251, 35000 Rennes, France
[107] INAF – Osservatorio Astronomico di Capodimonte, Via Moiariello 16, 80131, Napoli, Italy
[108] Niels Bohr Institute, University of Copenhagen, Juliane Maries Vej 30, 2100 Copenhagen Ø, Denmark
[109] Las Cumbres Observatory, 6740 Cortona Drive Suite 102, Goleta, CA 93117, USA
[110] Astrophysics Research Institute, Liverpool John Moores University, 146 Brownlow Hill, Liverpool L3 5RF, UK
[111] IPAC, Mail Code 100-22, California Institute of Technology, 1200 E. California Blvd., Pasadena, CA 91125, USA
[112] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, M/S 169-327, Pasadena, CA 91109, USA
[113] IRAP, Université de Toulouse, CNRS, UPS, CNES, 9 Av. colonel Roche, BP 44346, 31028 Toulouse Cedex 4, France
[114] Konkoly Observatory, Research Centre for Astronomy and Earth Sciences, MTA Centre of Excellence, Konkoly Thege Miklós út 15-17, 1121 Budapest, Hungary
[115] MTA CSFK Lendület Near-Field Cosmology Research Group, Konkoly Observatory, CSFK, Konkoly Thege Miklós út 15-17, 1121 Budapest, Hungary
[116] ELTE Eötvös Loránd University, Institute of Physics, 1117, Pázmány Péter sétány 1A, Budapest, Hungary
[117] Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia
[118] Institute of Theoretical Physics, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic
[119] INAF – Osservatorio Astronomico di Brera, Via E. Bianchi 46, 23807 Merate (LC), Italy
[120] AKKA for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[121] Departmento de Física de la Tierra y Astrofísica, Universidad Complutense de Madrid, 28040 Madrid, Spain
[122] Vitrociset Belgium for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[123] Department of Astrophysical Sciences, 4 Ivy Lane, Princeton University, Princeton NJ 08544, USA
[124] Departamento de Astrofísica, Centro de Astrobiología (CSIC-INTA), ESA-ESAC. Camino Bajo del Castillo s/n, 28692 Villanueva de la Cañada, Madrid, Spain
[125] naXys, University of Namur, Rempart de la Vierge, 5000 Namur, Belgium
[126] EPFL – Ecole Polytechnique fédérale de Lausanne, Institute of Mathematics, Station 8 EPFL SB MATH SDS, Lausanne, Switzerland
[127] H H Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol BS8 1TL, UK
[128] Sorbonne Université, CNRS, UMR7095, Institut d'Astrophysique de Paris, 98bis bd. Arago, 75014 Paris, France
[129] Porter School of the Environment and Earth Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[130] Laboratoire Univers et Particules de Montpellier, Université Montpellier, Place Eugène Bataillon, CC72, 34095 Montpellier Cedex 05, France
[131] Faculty of Mathematics and Physics, University of Ljubljana, Jadranska ulica 19, 1000 Ljubljana, Slovenia

Appendix A: Details of the random forest classifier parameters and training set.

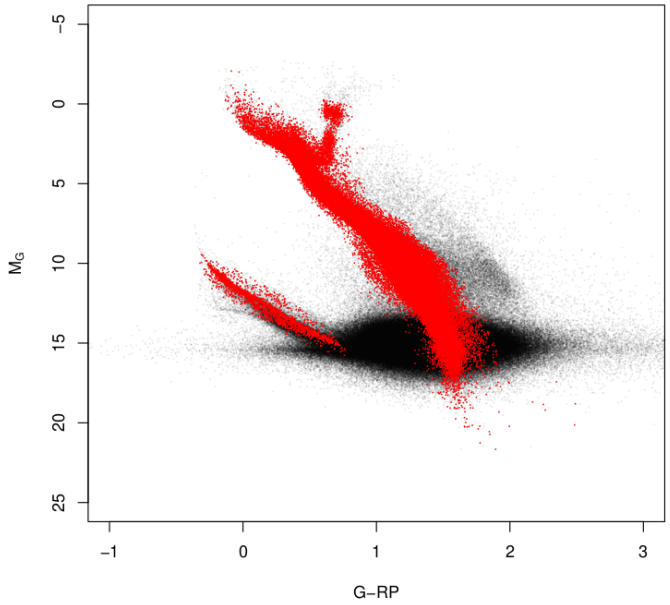

Fig. A.1. CAMD with colour G–$G_{\rm RP}$ and absolute magnitude M_G of the set of sources with parallaxes greater than or equal to 8 mas (black) and those used as examples of good astrometry (red points) in the random forest training set.

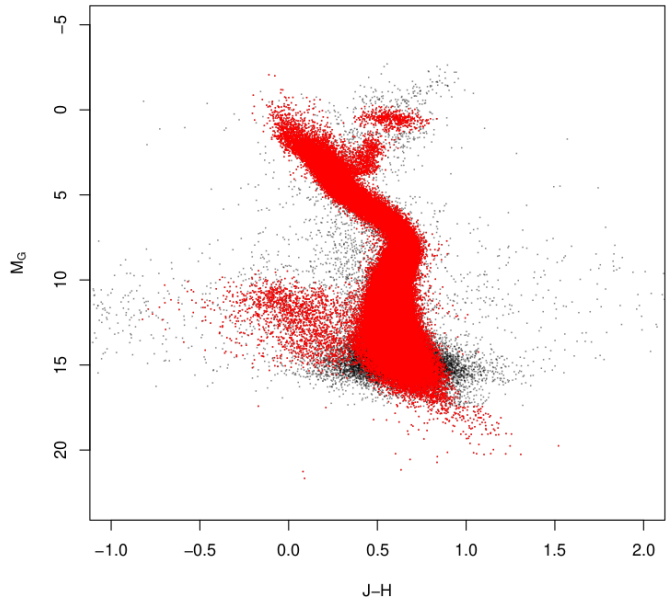

Fig. A.2. As Fig. A.1 for M_G and the 2MASS colour index J–H.

A.1. Colour-absolute magnitude diagrams

Figures A.1–A.3 show the position in several colour-absolute magnitude diagrams of the sources selected as examples of good astrometry in the training set (red) superimposed on the full distribution of sources with observed parallaxes greater than or equal to 8 mas. Figures A.3 and A.2 show that the requirement to have a 2MASS counterpart to the *Gaia* source already removes most of the sources with spurious observed parallaxes greater than 8 mas.

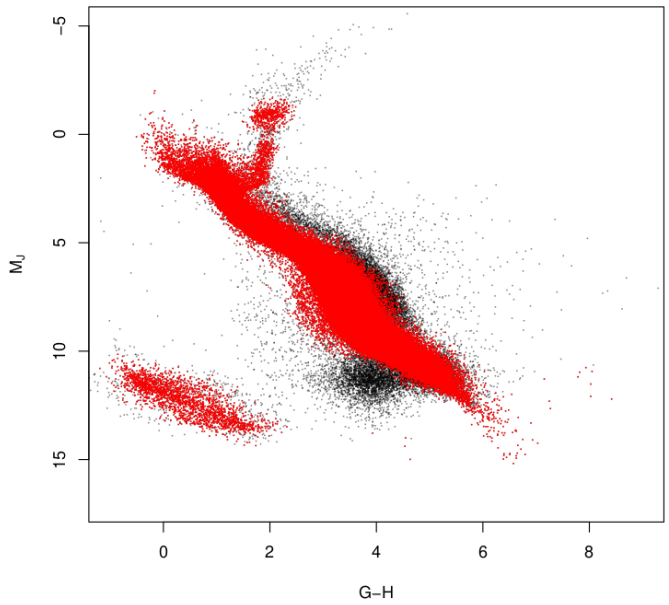

Fig. A.3. As Fig. A.1 for M_J and the G–H colour index.

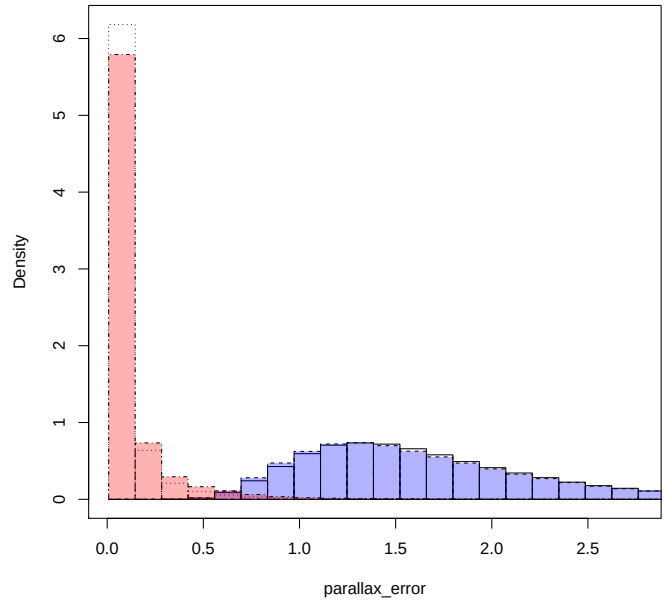

Fig. A.4. Distribution of values of the parallax_error feature in the set of training examples of the bad category (continuous line, white filling); in the set of training examples of the good category (dotted line, white filling); the set of sources classified as bad astrometric solutions (dashed line, blue transparent filling); and the set of sources classified as good astrometric solutions (dash-dotted line, red transparent filling).

A.2. Parameters tested for relevance

Table A.1 lists all catalogue columns tested for relevance in the classification problem of separating good astrometric solutions from spurious ones. We did not check for the relevance of astrometric_primary_flag, astrometric_weight_ac", nu_eff_used_in_astrometry, pseudocolour, and pseudocolour_error due to the high fraction of missing values in the training set.

A.3. Distributions of features

Table A.1. Importance of all features tested for classification by the Random Forest classifier ordered according to the mean decrease in accuracy (two leftmost columns) and by the mean decrease in the Gini index (two rightmost columns).

Feature name	Mean decrease accuracy	Feature name	Mean decrease Gini index
parallax_error	0.125	parallax_error	33821
parallax_over_error	0.087	parallax_over_error	27713
pmra	0.056	astrometric_sigma5d_max	24035
astrometric_sigma5d_max	0.052	pmra_error	20226
pmdec	0.047	pmdec_error	14866
pmdec_error	0.027	astrometric_excess_noise	12737
pmra_error	0.025	astrometric_params_solved	7677
astrometric_excess_noise	0.013	ipd_gof_harmonic_amplitude	5628
visibility_periods_used	0.01	ruwe	3383
ruwe	0.008	visibility_periods_used	2371
astrometric_gof_al	0.005	pmdec	2263
astrometric_n_obs_ac	0.005	pmra	2039
ipd_gof_harmonic_amplitude	0.004	ipd_frac_odd_win	1566
astrometric_excess_noise_sig	0.003	ipd_frac_multi_peak	1006
ipd_frac_odd_win	0.002	astrometric_gof_al	801
astrometric_chi2_al	0.002	scan_direction_strength_k2	694
parallax_pmdec_corr	0.002	parallax_pmdec_corr	522
ipd_frac_multi_peak	0.002	astrometric_excess_noise_sig	413
scan_direction_strength_k2	0.001	astrometric_n_good_obs_al	394
astrometric_n_good_obs_al	0.001	astrometric_chi2_al	275
astrometric_params_solved	0.001	astrometric_n_obs_al	244
astrometric_n_obs_al	0.001	astrometric_n_obs_ac	224
astrometric_matched_transits	0.001	dec_parallax_corr	208
dec_parallax_corr	0.001	astrometric_matched_transits	165
dec_pmdec_corr	0.001	dec_pmdec_corr	157
scan_direction_mean_k2	0.001	ra_dec_corr	65
ra_parallax_corr	0	scan_direction_strength_k1	59
scan_direction_strength_k4	0	scan_direction_mean_k2	50
ra_dec_corr	0	scan_direction_strength_k4	50
scan_direction_strength_k1	0	parallax_pmra_corr	49
scan_direction_mean_k4	0	ra_parallax_corr	48
scan_direction_strength_k3	0	ra_pmdec_corr	44
parallax_pmra_corr	0	scan_direction_mean_k4	42
astrometric_n_bad_obs_al	0	scan_direction_strength_k3	41
ra_pmdec_corr	0	astrometric_n_bad_obs_al	38
scan_direction_mean_k3	0	scan_direction_mean_k3	30
ipd_gof_harmonic_phase	0	ipd_gof_harmonic_phase	29
pmra_pmdec_corr	0	ra_pmra_corr	28
scan_direction_mean_k1	0	pmra_pmdec_corr	27
ra_pmra_corr	0	scan_direction_mean_k1	24
dec_pmra_corr	0	dec_pmra_corr	22

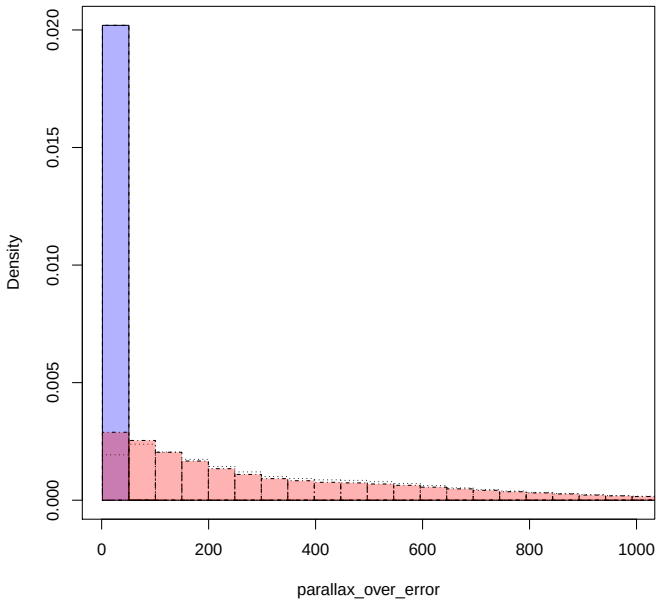

Fig. A.5. Distribution of values of the parallax_over_error feature in the set of training examples of the bad category (continuous line, white filling); in the set of training examples of the good category (dotted line, white filling); the set of sources classified as bad astrometric solutions (dashed line, blue transparent filling); and the set of sources classified as good astrometric solutions (dash-dotted line, red transparent filling).

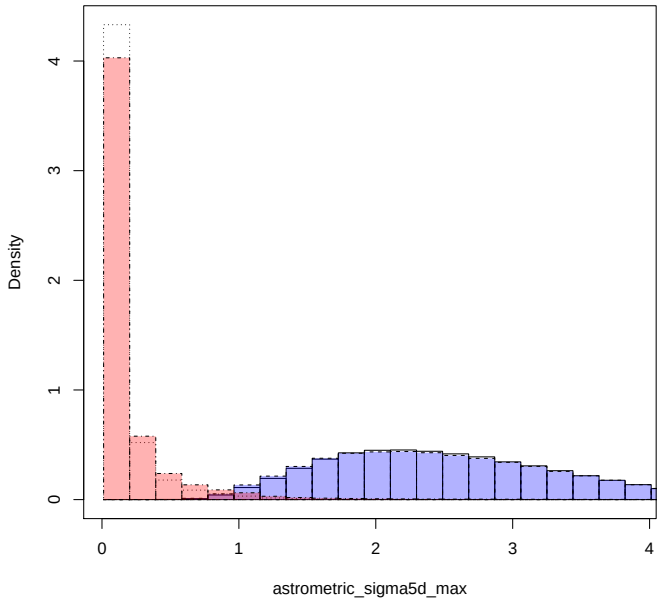

Fig. A.6. Distribution of values of the astrometric_sigma5d_max feature in the set of training examples of the bad category (continuous line, white filling); in the set of training examples of the good category (dotted line, white filling); the set of sources classified as bad astrometric solutions (dashed line, blue transparent filling); and the set of sources classified as good astrometric solutions (dash-dotted line, red transparent filling).

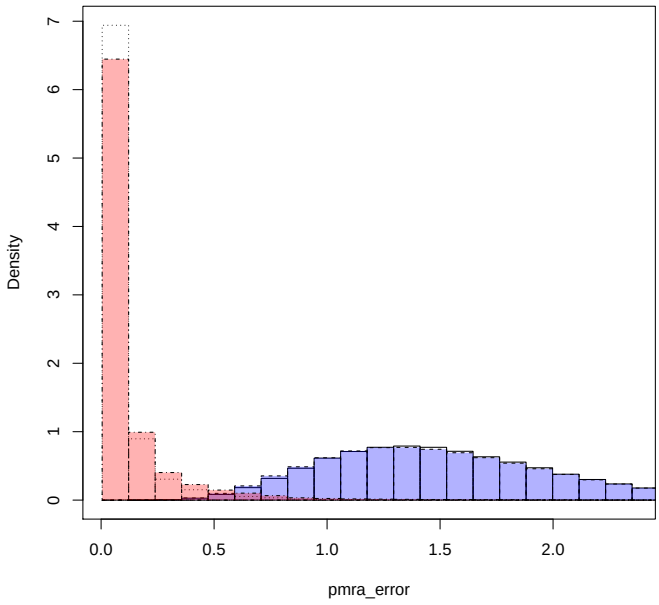

Fig. A.7. Distribution of values of the pmra_error feature in the set of training examples of the bad category (continuous line, white filling); in the set of training examples of the good category (dotted line, white filling); the set of sources classified as bad astrometric solutions (dashed line, blue transparent filling); and the set of sources classified as good astrometric solutions (dash-dotted line, red transparent filling).

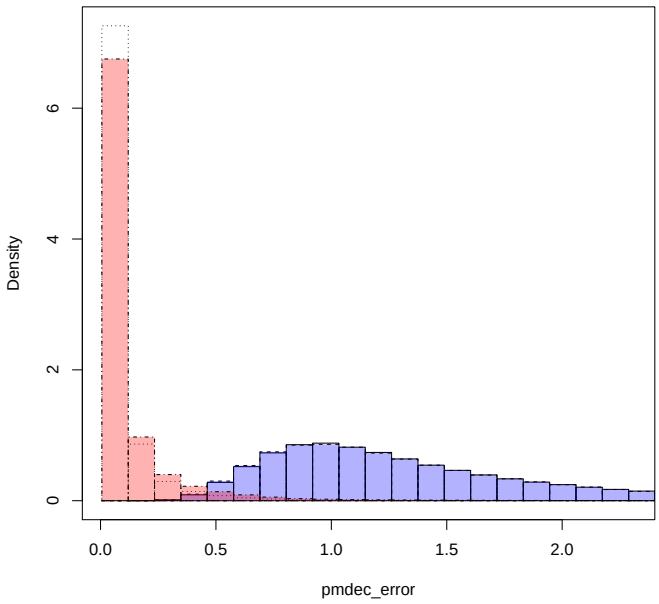

Fig. A.8. Distribution of values of the pmdec_error feature in the set of training examples of the bad category (continuous line, white filling); in the set of training examples of the good category (dotted line, white filling); the set of sources classified as bad astrometric solutions (dashed line, blue transparent filling); and the set of sources classified as good astrometric solutions (dash-dotted line, red transparent filling).

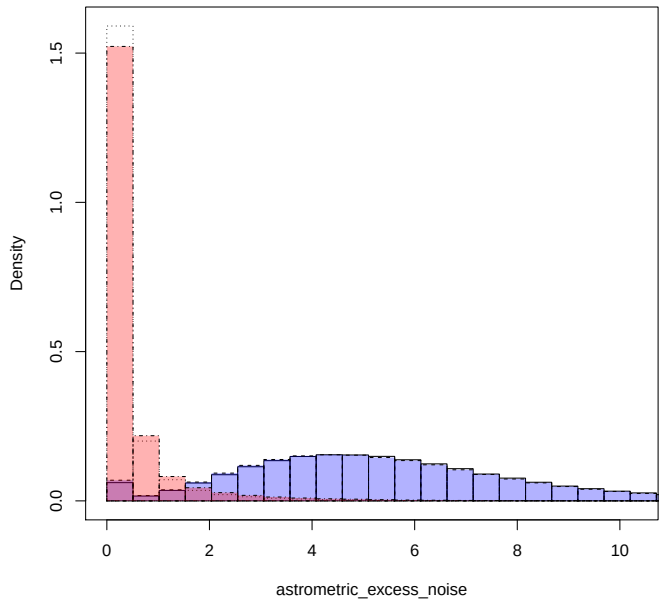

Fig. A.9. Distribution of values of the astrometric_excess_noise feature in the set of training examples of the bad category (continuous line, white filling); in the set of training examples of the good category (dotted line, white filling); the set of sources classified as bad astrometric solutions (dashed line, blue transparent filling); and the set of sources classified as good astrometric solutions (dash-dotted line, red transparent filling).

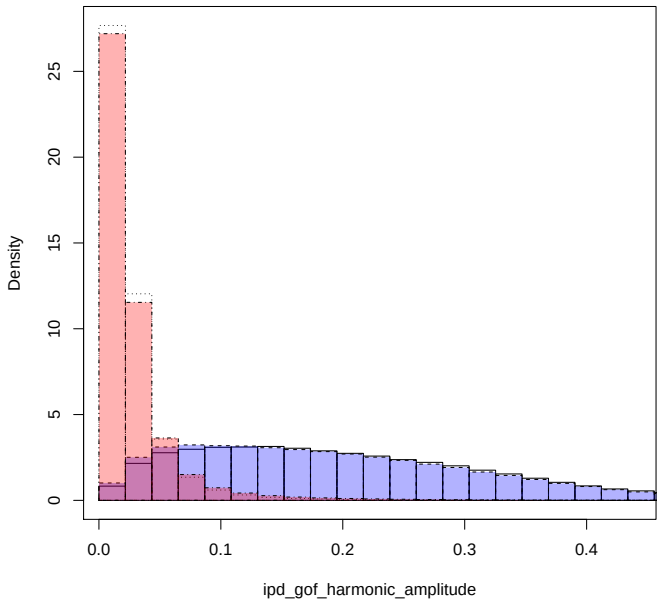

Fig. A.10. Distribution of values of the ipd_gof_harmonic_amplitude feature in the set of training examples of the bad category (continuous line, white filling); in the set of training examples of the good category (dotted line, white filling); the set of sources classified as bad astrometric solutions (dashed line, blue transparent filling); and the set of sources classified as good astrometric solutions (dash-dotted line, red transparent filling).

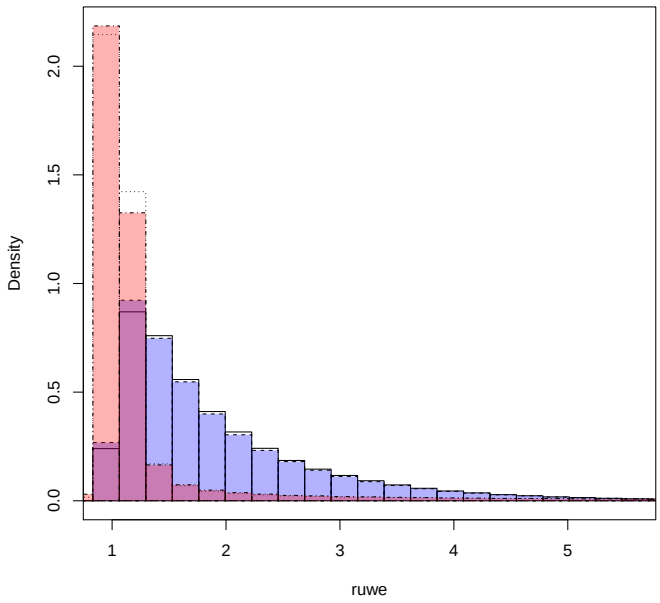

Fig. A.11. Distribution of values of the ruwe feature in the set of training examples of the bad category (continuous line, white filling); in the set of training examples of the good category (dotted line, white filling); the set of sources classified as bad astrometric solutions (dashed line, blue transparent filling); and the set of sources classified as good astrometric solutions (dash-dotted line, red transparent filling).

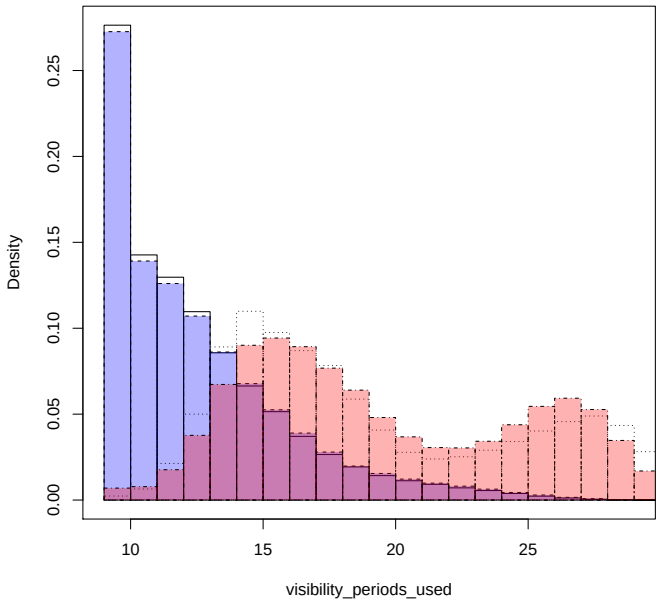

Fig. A.12. Distribution of values of the visibility_periods_used feature in the set of training examples of the bad category (continuous line, white filling); in the set of training examples of the good category (dotted line, white filling); the set of sources classified as bad astrometric solutions (dashed line, blue transparent filling); and the set of sources classified as good astrometric solutions (dash-dotted line, red transparent filling).

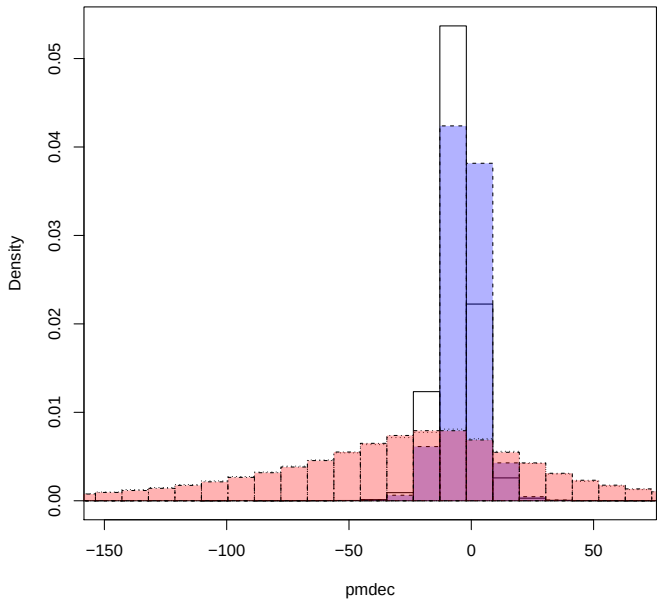

Fig. A.13. Distribution of values of the pmdec feature in the set of training examples of the bad category (continuous line, white filling); in the set of training examples of the good category (dotted line, white filling); the set of sources classified as bad astrometric solutions (dashed line, blue transparent filling); and the set of sources classified as good astrometric solutions (dash-dotted line, red transparent filling).

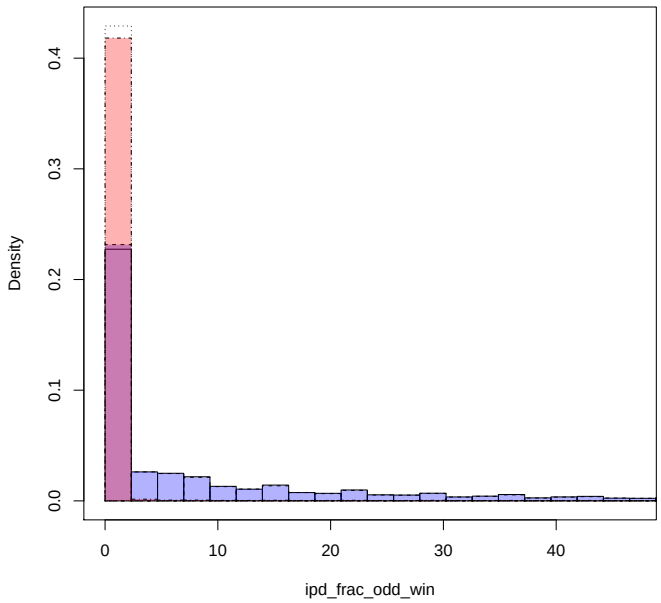

Fig. A.15. Distribution of values of the ipd_frac_odd_win feature in the set of training examples of the bad category (continuous line, white filling); in the set of training examples of the good category (dotted line, white filling); the set of sources classified as bad astrometric solutions (dashed line, blue transparent filling); and the set of sources classified as good astrometric solutions (dash-dotted line, red transparent filling).

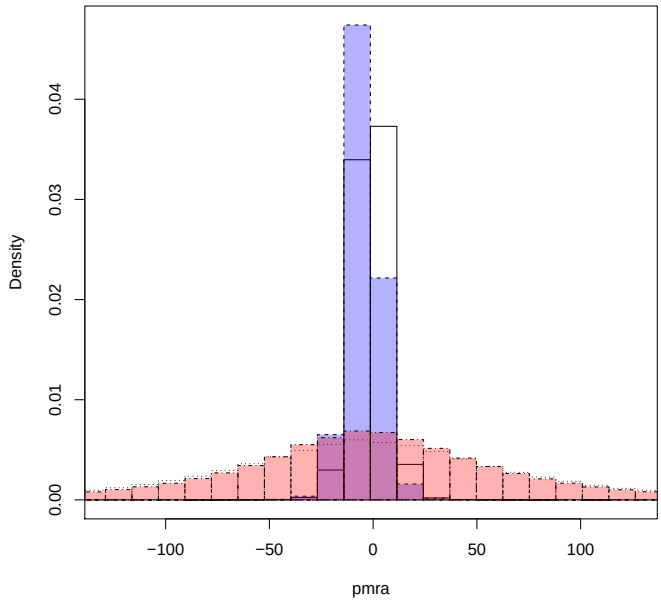

Fig. A.14. Distribution of values of the pmra feature in the set of training examples of the bad category (continuous line, white filling); in the set of training examples of the good category (dotted line, white filling); the set of sources classified as bad astrometric solutions (dashed line, blue transparent filling); and the set of sources classified as good astrometric solutions (dash-dotted line, red transparent filling).

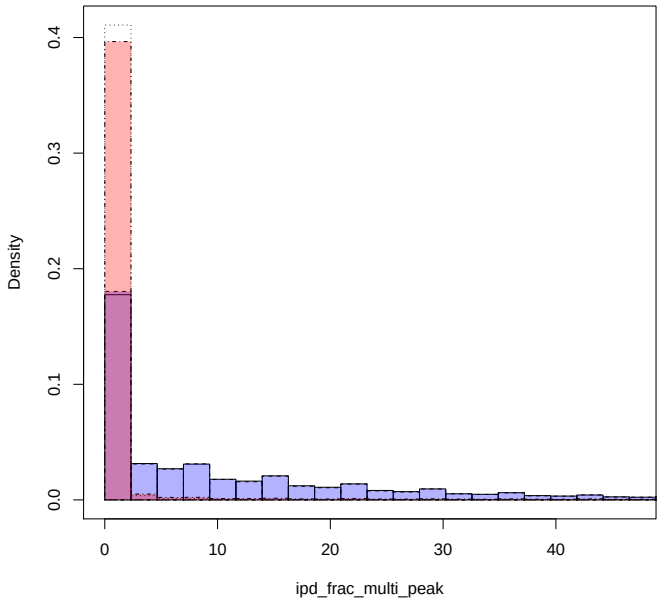

Fig. A.16. Distribution of values of the ipd_frac_multi_peak feature in the set of training examples of the bad category (continuous line, white filling); in the set of training examples of the good category (dotted line, white filling); the set of sources classified as bad astrometric solutions (dashed line, blue transparent filling); and the set of sources classified as good astrometric solutions (dash-dotted line, red transparent filling).

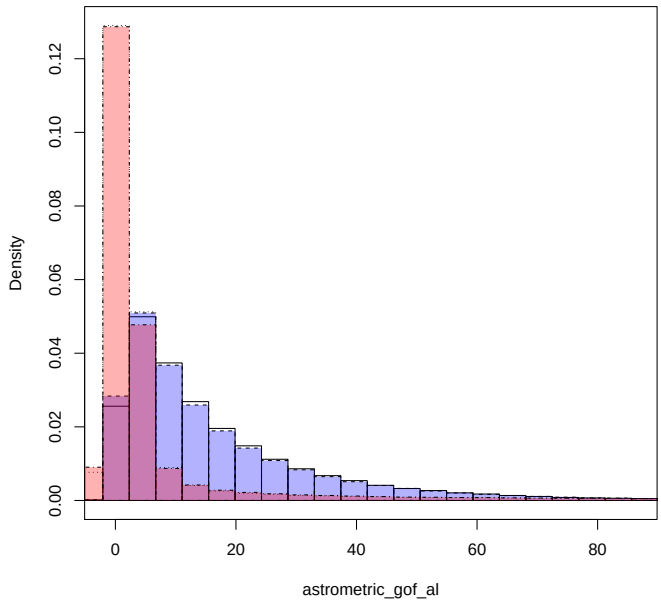

Fig. A.17. Distribution of values of the astrometric_gof_al feature in the set of training examples of the bad category (continuous line, white filling); in the set of training examples of the good category (dotted line, white filling); the set of sources classified as bad astrometric solutions (dashed line, blue transparent filling); and the set of sources classified as good astrometric solutions (dash-dotted line, red transparent filling).

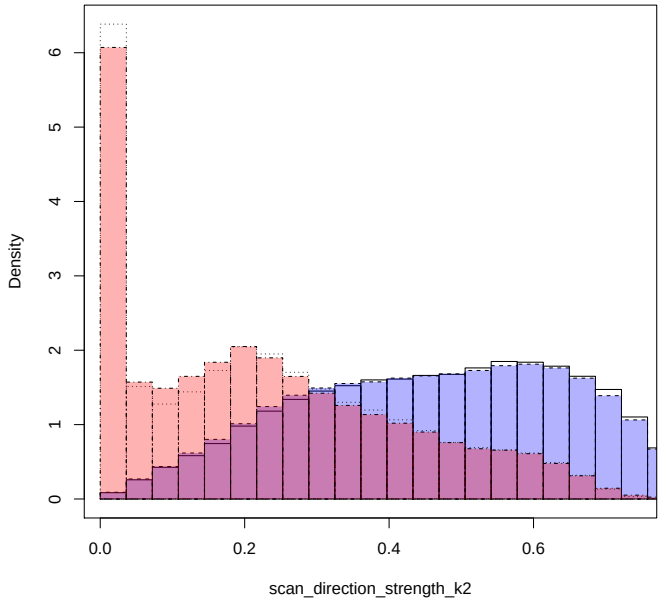

Fig. A.18. Distribution of values of the scan_direction_strength_k2 feature in the set of training examples of the bad category (continuous line, white filling); in the set of training examples of the good category (dotted line, white filling); the set of sources classified as bad astrometric solutions (dashed line, blue transparent filling); and the set of sources classified as good astrometric solutions (dash-dotted line, red transparent filling).

Appendix B: The relations and Gaussian Mixture Model priors used for the determination of space velocities in Galactic coordinates

The relations that define space velocities in terms of the observables, *Gaia* coordinates, parallaxes and proper motions, and radial velocities are:

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = A_{\mathrm{G}}^{T} \cdot A \begin{pmatrix} 4.74047\,\mu_{\alpha^*}/\varpi \\ 4.74047\,\mu_{\delta}/\varpi \\ v_r \end{pmatrix} \tag{B.1}$$

where A_{G} is the transformation matrix to Galactic coordinates from the introduction to the HIPPARCOS catalogue (ESA 1997) and matrix A is obtained from the components of the normal triad at the star as:

$$A = \begin{pmatrix} -\sin\alpha & -\sin\delta\cos\alpha & \cos\delta\cos\alpha \\ \cos\alpha & -\sin\delta\sin\alpha & \cos\delta\sin\alpha \\ 0 & \cos\delta & \sin\delta \end{pmatrix} \tag{B.2}$$

The Bayesian model used to infer posterior probabilities for the space velocities requires the definition of priors for the model parameters. As described in Sect. 3, we fit Gaussian Mixture Models to a local (140 pc) simulation from the Besançon Galaxy model (Robin et al. 2003) and modify the result by adding a wide non-informative component. The resulting priors used in the inference process are defined in Eqs. (B.3)–(B.5) using the notation $\mathcal{N}(\cdot \mid \mu, \sigma)$ to denote the Gaussian distribution centred at μ and with standard deviation σ.

$$\begin{aligned} \pi(U) = {} & 0.52 \cdot \mathcal{N}(U \mid -11.3, 23.2) \\ & + 0.45 \cdot \mathcal{N}(U \mid -11, 44) \\ & + 0.03 \cdot \mathcal{N}(U \mid 0, 120) \end{aligned} \tag{B.3}$$

$$\begin{aligned} \pi(V) = {} & 0.588 \cdot \mathcal{N}(V \mid -26.1, 23.7) \\ & + 0.375 \cdot \mathcal{N}(V \mid -13, 11.3) \\ & + 0.03 \cdot \mathcal{N}(V \mid 0, 120) \\ & + 0.007 \cdot \mathcal{N}(V \mid -115.8, 114.3) \end{aligned} \tag{B.4}$$

$$\begin{aligned} \pi(W) = {} & 0.53 \cdot \mathcal{N}(W \mid -7.3, 19.4) \\ & + 0.2 \cdot \mathcal{N}(W \mid -10, 9.2) \\ & + 0.21 \cdot \mathcal{N}(W \mid -4.1, 10.1) \\ & + 0.03 \cdot \mathcal{N}(W \mid -7, 43.3) \\ & + 0.03 \cdot \mathcal{N}(W \mid 0, 120) \end{aligned} \tag{B.5}$$